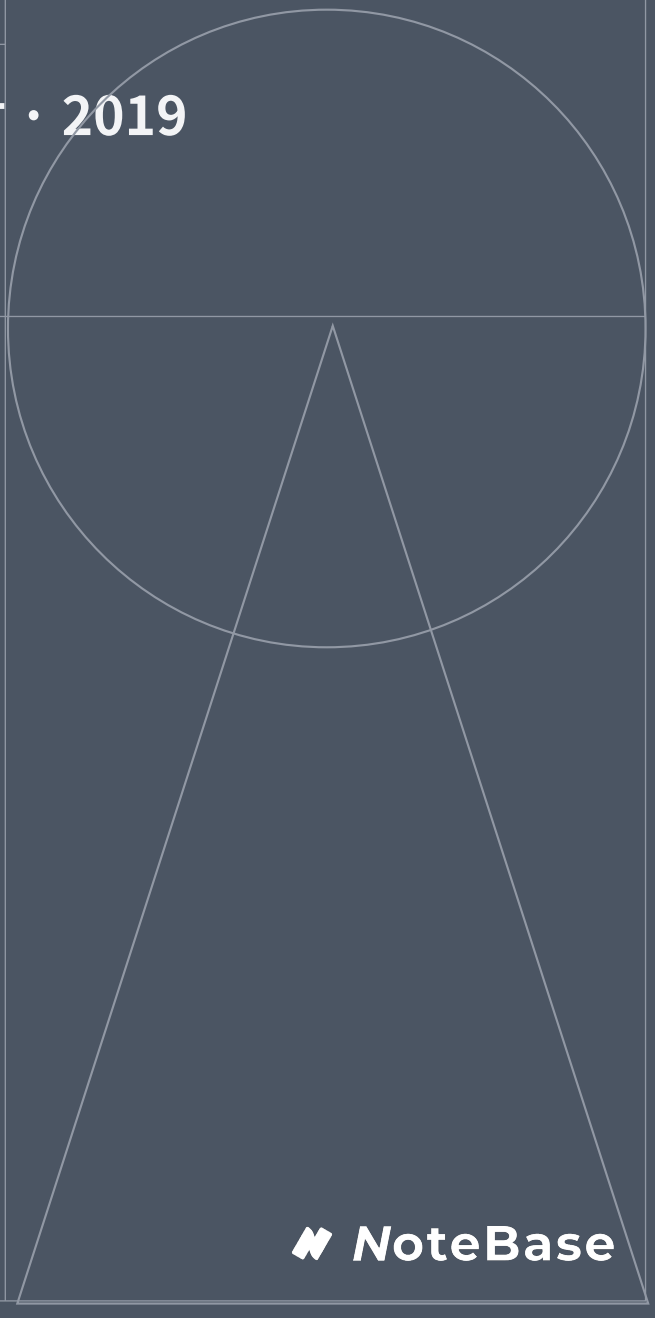
기출OX 해설집

공무원/국가직/7급 · 통계학 · 2019



NoteBase

기출OX 해설집

공무원/국가직/7급 · 통계학 · 2019

공무원/국가직/7급 · 통계학 · 2019

문 1. 다음은 어느 회사의 방문 고객 수에 대한 자료이다. 산술평균, 중앙값, 최빈값을 모두 더한 값은?

1, 2, 3, 3, 2, 3, 4, 6

① 7

② 8

③ 9

④ 10

정답 3번

출제 포인트

이 문제는 주어진 자료에서 산술평균, 중앙값, 최빈값을 각각 구한 뒤 그 합을 계산하는 문제이다.

조건 정리

주어진 자료: 1, 2, 3, 3, 2, 3, 4, 6
자료 개수: 8개
구해야 할 값: 산술평균 + 중앙값 + 최빈값

사용할 공식

산술평균은 자료의 합자료의 개수이다.
중앙값은 자료를 크기순으로 배열했을 때 가운데 값이며, 자료 수가 짝수이면 가운데 두 값의 평균이다.
최빈값은 가장 많이 나타나는 값이다.
이 문제는 대표값 세 가지를 모두 묻고 있으므로 각 정의를 그대로 적용하면 된다.

NoteBase

계산 과정

1. 자료의 합은 $1 + 2 + 3 + 3 + 2 + 3 + 4 + 6 = 24$이므로 산술평균은 $24 \div 8 = 3$이다.
2. 자료를 크기순으로 정리하면 1, 2, 2, 3, 3, 3, 4, 6이다. 8개 자료의 가운데 두 값은 4번째와 5번째 값이므로 중앙값은 $\frac{3+3}{2} = 3$이다.
3. 가장 많이 나온 값은 3으로 3번 나타나므로 최빈값은 3이다.
4. 따라서 산술평균, 중앙값, 최빈값의 합은 $3 + 3 + 3 = 9$이다.

선지별 해설

① X: 7은 계산 결과와 다르다. 틀린 표현은 '세 대표값의 합이 7이다'이다. 올바르게는 '세 대표값의 합이 9이다'이다.
② X: 8은 계산 결과와 다르다. 틀린 표현은 '세 대표값의 합이 8이다'이다. 올바르게는 '세 대표값의 합이 9이다'이다. 이 값은 중앙값이나 최빈값 중 하나를 잘못 읽었을 때 나올 수 있다.
③ O: 산술평균 3, 중앙값 3, 최빈값 3이므로 합은 9이다.
④ X: 10은 계산 결과와 다르다. 틀린 표현은 '세 대표값의 합이 10이다'이다. 올바르게는 '세 대표값의 합이 9이다'이다. 이 값은 자료의 개수나 중앙값 계산을 잘못 처리했을 때 고르기 쉽다.

오답 포인트

자료를 정렬하지 않은 채 중앙값을 바로 읽거나, 짝수 개 자료의 중앙값을 가운데 한 개 값으로 착각하면 오답이 된다. 최빈값도 가장 큰 값이 아니라 가장 자주 나온 값이라는 점을 혼동하기 쉽다.

다시 나오면 이렇게 판단

대표값 문제는 먼저 자료를 정렬하고, 산술평균은 합과 개수로 계산하며, 중앙값은 자료 개수의 홀짝을 확인한 뒤, 최빈값은 빈도를 세어 판단하면 된다.

암기 포인트

산술평균은 합 나누기 개수, 중앙값은 가운데 값, 최빈값은 가장 자주 나온 값이다.

NoteBase

문 2. 다음 사례에 해당하는 표본추출 방법은?

어느 여론조사 기관이 전국의 유권자 중 무작위로 뽑힌 1,000명에게 차기 대통령 후보에 대한 선호도를 물었다. 이때, 표본의 40%가 A 후보를, 35%가 B 후보를, 15%가 C 후보를 지지하는 결과를 얻었다.

① 단순임의추출법(simple random sampling)
② 군집추출법(cluster sampling)
③ 층화추출법(stratified sampling)
④ 계통추출법(systematic sampling)

정답 1번

출제 포인트

이 문제는 주어진 사례에 맞는 표본추출 방법을 구별할 수 있는지를 묻는 문제이다.

사례 정리

모집단은 전국의 유권자이다.
전국 유권자 중에서 1,000명을 무작위로 뽑아 조사하였다.
지역별, 연령별 등으로 층을 나누었다는 정보도 없고, 일정 간격으로 뽑거나 특정 집단을 통째로 뽑았다는 정보도 없다.

핵심 개념

단순임의추출법은 모집단 전체에서 각 구성원이 뽑힐 기회를 같게 하여 무작위로 표본을 뽑는 방법이다.
적용 요건은 전체 모집단에서 직접 무작위 추출이 이루어지고, 층화나 군집화, 일정한 간격 추출 같은 추가 설계가 없어야 한다.

적용 기준

이 사례에서는 '전국의 유권자 중 무작위로 뽑은 1,000명'이라는 표현이 핵심이다. 이는 모집단 전체에서 개별 유권자를 직접 무작위로 뽑은 것이므로 단순임의추출법에 해당한다.

NoteBase

선지별 해설

1번 O: 전국 유권자 전체를 모집단으로 두고 그중 1,000명을 무작위로 뽑았으므로 단순임의추출법이다.
2번 X: 군집추출법은 지역, 학교, 가구처럼 자연스럽게 묶인 집단을 먼저 뽑고 그 집단을 조사하는 방법이다. 이 사례에서 틀린 표현은 '집단을 먼저 뽑는다'는 전제이다. 올바르게는 전국 유권자 개개인 중에서 직접 무작위로 뽑았으므로 단순임의추출법이다.
3번 X: 층화추출법은 성별, 연령, 지역 등 동질적인 층으로 나눈 뒤 각 층에서 표본을 뽑는 방법이다. 이 사례에서 틀린 표현은 '층을 나누어 추출한다'는 부분이다. 올바르게는 별도의 층 구분 없이 전체 모집단에서 바로 무작위 추출하였다.
4번 X: 계통추출법은 명부를 일정 순서로 배열한 뒤 매 k번째처럼 일정 간격으로 표본을 뽑는 방법이다. 이 사례에서 틀린 표현은 '일정 간격으로 뽑는다'는 부분이다. 올바르게는 간격 규칙 없이 무작위로 1,000명을 뽑았으므로 계통추출법이 아니다.

오답 포인트

문항에 전국 단위 조사라는 말이 나오면 지역별 층화추출이나 군집추출로 오해하기 쉽다. 그러나 전국 단위라는 말은 모집단의 범위를 뜻할 뿐이고, 실제 추출 절차에 층 구분이나 집단 선정이 제시되지 않았다는 점이 핵심이다.

다시 나오면 이렇게 판단

표본추출 문제는 먼저 개별 원소를 직접 뽑는지, 집단을 먼저 뽑는지, 층을 나누는지, 일정 간격을 쓰는지를 순서대로 확인하면 된다. 아무 추가 절차 없이 전체에서 무작위로 뽑으면 단순임의추출법이다.

암기 포인트

전체에서 바로 무작위로 뽑으면 단순임의, 층을 나누면 층화, 집단을 뽑으면 군집, 간격을 두면 계통이다.

NoteBase

문 3. 확률질량함수 또는 확률밀도함수가 될 수 없는 것은?

① $f(x) = \frac{1}{10}$, ($0 < x \leq 10$인 정수)

② $f(x) = \left(\frac{1}{2}\right)^x$, ($x$는 자연수)

③ $f(x) = e^{-x}$, ($x > 0$)

④ $f(x) = \frac{1}{3}(3 - 2x)$, ($0 \leq x \leq 1$)

정답 4번

출제 포인트

이 문제는 확률질량함수와 확률밀도함수의 공통 조건인 음이 아닌 값과 전체 합 또는 전체 적분값이 1인지 판단하는 문제이다.

핵심 개념

확률질량함수는 이산형 확률변수에 대해 각 값의 확률을 주는 함수이며, 모든 값이 0 이상이고 전체 합이 1이어야 한다.
확률밀도함수는 연속형 확률변수에 대해 구간확률을 적분으로 계산하는 함수이며, 함수값이 0 이상이고 정의역 전체에서의 적분값이 1이어야 한다.
판단 축은 이산형이면 합, 연속형이면 적분이라는 점이다.

선지별 해설

① O: $x = 1, 2, \ldots, 10$에서 각 값의 확률이 모두 $\frac{1}{10}$이므로 전체 합은 $10 \times \frac{1}{10} = 1$이다. 이산형에서 확률질량함수가 될 수 있다.
② O: x가 자연수일 때 Σ _x=1^∞(12)^x=(12)/(1-12)=1이다. 모든 값이 0 이상이므로 확률질량함수가 될 수 있다.
③ O: $x > 0$에서 $e^{-x} > 0$이고 $\int_0^\infty e^{-x}dx = 1$이다. 따라서 확률밀도함수가 될 수 있다.
④ X: 0≤ x≤1에서 13(3-2x)는 음수는 아니지만 ∫ _0^113(3-2x) dx=23이므로 전체 적분값이 1이 아니다. 틀린 부분은 계수 13이며, 올바르게는 f(x)=12(3-2x)처럼 전체 적분값이 1이 되도록 정규화되어야 한다.

오답 포인트

④는 구간에서 음수가 아니므로 얼핏 확률밀도함수처럼 보이지만, 확률함수 판단은 음이 아닌 조건만으로 끝나지 않는다. 특히 연속형에서는 전체 적분값이 1인지까지 반드시 확인해야 한다는 점을 놓치기 쉽다.

다시 나오면 이렇게 판단

먼저 정의역이 이산형인지 연속형인지 구분한다. 이산형이면 가능한 값들에서 확률의 합을 계산하고, 연속형이면 정의역 전체에서 적분값을 계산한다. 그다음 음이 아닌 조건까지 함께 확인하면 된다.

NoteBase

암기 포인트

이산형은 전체 합이 1, 연속형은 전체 적분값이 1이라는 점이 확률함수 판단의 핵심이다.

NoteBase

문 4. 다음은 표본으로 조사된 5명의 양팔 길이(X)와 키(Y)에 대한 자료 $(x_i, y_i), i = 1, \cdots, 5$이다. $\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = 23.6$일 때, 양팔 길이와 키의 표본상관계수는? (단, 단위는 cm이다)

구분	1	2	3	4	5	평균	표준편차
양팔 길이(X)	168	165	168	169	162	166.4	2.88
키(Y)	169	165	167	167	163	166.2	2.28

① $\frac{4 \times 2.88 \times 2.28}{23.6}$

② $\frac{23.6}{4 \times 2.88 \times 2.28}$

③ $\frac{5 \times 2.88 \times 2.28}{23.6}$

④ $\frac{23.6}{5 \times 2.88 \times 2.28}$

정답 2번

출제 포인트

이 문제는 표본상관계수 r를 편차곱의 합과 표본표준편차로 계산하는 문제이다.

조건 정리

주어진 값: 표본 크기 $n = 5$, $\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = 23.6$, $s_X = 2.88$, $s_Y = 2.28$이다.
구해야 할 값: 양팔 길이와 키의 표본상관계수 r이다.
주의할 조건: 표본표준편차가 주어졌으므로 분모에는 n이 아니라 $n - 1 = 4$가 들어간다.

사용할 공식

표본상관계수는 $r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y}$이다.
이 공식을 쓰는 이유: 문제에서 편차곱의 합과 두 변수의 표본표준편차가 함께 주어졌으므로 표본공분산을 거쳐 바로 표본상관계수를 구할 수 있다.

계산 과정

1. 표본공분산은 $s_{XY} = \frac{23.6}{5-1} = \frac{23.6}{4} = 5.9$이다.
2. 두 표본표준편차의 곱은 $2.88 \times 2.28 = 6.5664$이다.
3. 따라서 $r = \frac{5.9}{6.5664} = \frac{23.6}{4 \times 2.88 \times 2.28} \approx 0.898$이다.

NoteBase

선지별 해설

① X: 표본상관계수는 $\frac{23.6}{4 \times 2.88 \times 2.28}$인데, 이 선지는 분자와 분모를 뒤집었다. 틀린 표현은 $\frac{4 \times 2.88 \times 2.28}{23.6}$이고, 올바르게는 $\frac{23.6}{4 \times 2.88 \times 2.28}$이다.
② O: 계산 결과가 정확히 $\frac{23.6}{4 \times 2.88 \times 2.28} \approx 0.898$이므로 표본상관계수를 바르게 나타낸 선지이다.
③ X: 틀린 표현은 분모와 분자를 뒤집은 점과 5를 쓴 점이다. 표본상관계수는 $n = 5$ 자체가 아니라 $n - 1 = 4$를 써서 $\frac{23.6}{4 \times 2.88 \times 2.28}$로 계산한다.
④ X: 분자와 분모의 위치는 맞지만 틀린 표현은 분모의 5이다. 표본표준편차를 사용했으므로 올바르게는 $\frac{23.6}{4 \times 2.88 \times 2.28}$이다.

오답 포인트

③, ④처럼 분모에 n을 넣으면 모집단 공식과 표본 공식을 혼동한 것이다. ①, ③처럼 식을 뒤집는 실수는 편차곱의 합이 보이면 그것을 그대로 분자에 둔다는 원칙을 놓쳤을 때 나온다.

다시 나오면 이렇게 판단

표본상관계수 문제는 먼저 주어진 표준편차가 표본인지 모집단인지 확인한 뒤, $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y}$ 형태로 바로 대입해 판단한다.

암기 포인트

표본상관계수는 편차곱의 합을 $(n - 1)$과 두 표본표준편차의 곱으로 나눈 값이다.

NoteBase

문 5. 어떤 확률변수 X의 확률분포표가 다음과 같다. 이 모집단에서 복원추출한 크기가 2인 표본으로부터 얻은 표본평균을 $\overline{X}$라고 하자. 표본평균 $\overline{X}$의 기댓값이 $E(\overline{X})=\frac{1}{4}$일 때, 표본평균 $\overline{X}$의 분산 $\mathrm{Var}(\overline{X})$은? (단, a는 상수)

X	-1	0	1	합계
$\mathrm{P}(X=x)$	$\frac{1}{2}-a$	a	$\frac{1}{2}$	1

① $\frac{9}{32}$
② $\frac{11}{32}$
③ $\frac{9}{16}$
④ $\frac{11}{16}$

정답 2번

출제 포인트

이 문제는 $E(\bar{X})=E(X)$와 $\mathrm{Var}(\bar{X})=\mathrm{Var}(X)/n$을 이용해 미지수 a와 표본평균의 분산을 구하는 문제이다.

조건 정리

주어진 값은 $P(X=-1)=1/2-a$, $P(X=0)=a$, $P(X=1)=1/2$, 복원추출, 표본크기 2, $E(\bar{X})=1/4$이다.
구해야 할 값은 $\mathrm{Var}(\bar{X})$이다.
복원추출이므로 두 표본은 독립이고, 표본평균의 분산은 원래 분산을 2로 나눈 값이 된다.

사용할 공식

$E(\bar{X})=E(X)$, $\mathrm{Var}(\bar{X})=\frac{\mathrm{Var}(X)}{2}$, $\mathrm{Var}(X)=E(X^2)-\{E(X)\}^2$
이 공식을 쓰는 이유는 복원추출한 크기 2의 표본평균이 서로 독립인 두 관측값의 평균이기 때문이다.

계산 과정

1. $E(X)=(-1)(1/2-a)+0\cdot a+1\cdot(1/2)=a$이므로 $E(\bar{X})=E(X)=1/4$에서 $a=1/4$이다.
2. $E(X^2)=(-1)^2(1/2-a)+0^2\cdot a+1^2\cdot(1/2)=1-a=3/4$이다.
3. 따라서 $\mathrm{Var}(X)=E(X^2)-\{E(X)\}^2=3/4-(1/4)^2=3/4-1/16=11/16$이다.
4. 그러므로 $\mathrm{Var}(\bar{X})=\frac{\mathrm{Var}(X)}{2}=\frac{11/16}{2}=11/32$이다.

NoteBase

선지별 해설

① X: 계산 결과는 9/32가 아니라 11/32이다. 이 값은 $a = 1/4$를 반영한 분산 계산이 어긋난 경우에 나오는 오답이다. 올바르게는 표본평균의 분산은 11/32이다.
② O: $a = 1/4$, $\mathrm{Var}(X) = 11/16$, $\mathrm{Var}(\bar{X}) = 11/32$로 이어지므로 계산 결과와 정확히 일치한다.
③ X: 9/16은 표본평균의 분산으로 볼 수 없고, 표본크기 2에 따른 나눗셈도 반영되지 않은 값이다. 올바르게는 표본평균의 분산은 11/32이다.
④ X: 11/16은 원래 확률변수 X의 분산 $\mathrm{Var}(X)$이고, 문제에서 묻는 $\mathrm{Var}(\bar{X})$는 여기에 2를 나눈 11/32이다. 올바르게는 표본평균의 분산은 11/32이다.

오답 포인트

가장 큰 함정은 ④처럼 원래 확률변수의 분산과 표본평균의 분산을 구분하지 못하는 점이다. 또 $E(\bar{X}) = E(X)$에서 바로 $a = 1/4$를 얻는 구조를 놓치면 이후 계산이 모두 어긋난다.

다시 나오면 이렇게 판단

이런 문제는 먼저 $E(\bar{X}) = E(X)$로 미지수를 정리하고, 다음으로 $E(X^2)$와 $\mathrm{Var}(X)$를 구한 뒤, 마지막에 표본크기 n을 적용해 $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n$으로 마무리하는 순서로 판단한다.

암기 포인트

복원추출한 표본평균은 기대값은 그대로이고 분산은 $1/n$배가 된다.

NoteBase

문 6. $c(c \geq 2)$**개의 범주로 나누어지는** $r(r \geq 2)$**개의 다항모집단에서** $i(i = 1, \cdots, r)$**번째 모집단의 각 범주에 대한 모비율은** $p_{i1}, p_{i2}, \cdots, p_{ic}$**이고, 관측값은** $O_{i1}, O_{i2}, \cdots, O_{ic}$**이고,** i**번째 모집단의 표본의 크기는** n_i**이고** $n = \sum_{i=1}^{r} n_i$**이며,** $j = 1, \cdots, c$**에 대해** $O_{\cdot j} = \sum_{i=1}^{r} O_{ij}$**이다.** $r \times c$ **분할표의 동질성검정에 대한 설명으로 옳지 않은 것은?**

① 모든 $j(j = 1, \cdots, c)$에 대해 $\sum_{i=1}^{r} p_{ij} = 1$이다.

② 귀무가설은 각 $j(j = 1, \cdots, c)$에 대해 $p_{1j} = p_{2j} = \cdots = p_{rj}$이다.

③ 귀무가설하에서 $j(j = 1, \cdots, c)$에 속할 공통인 각 범주의 모비율의 추정값은 $\frac{O_{\cdot j}}{n}$이다.

④ 귀무가설하에서 i번째 모집단의 j번째 범주의 추정 기대도수는 $\frac{n_i O_{\cdot j}}{n}$이다.

정답 1번

출제 포인트

이 문제는 $r \times c$ 분할표의 동질성검정에서 귀무가설, 공통 모비율의 추정, 기대도수 공식을 정확히 구분하는지를 묻는 문제이다.

핵심 개념

동질성검정은 여러 모집단이 같은 범주분포를 가지는지를 검정하는 절차이다.
각 모집단 i의 범주별 모비율은 $p_{i1}, p_{i2}, \ldots, p_{ic}$이고, 한 모집단 안에서는 범주 전체의 합이 1이므로 $\sum_{j=1}^{c} p_{ij} = 1$이다.
귀무가설이 성립하면 각 범주 j의 공통 모비율은 표본을 합쳐 $\hat{p}_j = O_{\cdot j}/n$으로 추정하고, 기대도수는 $E_{ij} = n_i O_{\cdot j}/n$이다.

NoteBase

선지별 해설

① X
틀린 부분은 $\sum_{j=1}^{r} p_{ij} = 1$이라는 표현이다.
p_{ij}는 i번째 모집단 안에서 j범주에 속할 확률이므로, 합이 1이 되는 방향은 i 고정 후 j에 대해 더하는 경우이다.
올바르게는 각 i에 대해 $\sum_{j=1}^{c} p_{ij} = 1$이다.

② O
동질성검정의 귀무가설은 모든 모집단의 범주별 분포가 같다는 뜻이다. 따라서 각 범주 j에 대해 $p_{1j} = p_{2j} = \cdots = p_{rj}$로 둔다.

③ O
귀무가설하에서는 모집단들이 동일한 범주비율을 공유하므로 각 범주 j의 공통 모비율은 전체 표본에서 그 범주에 들어간 비율로 추정한다. 따라서 추정값은 $O_{\cdot j}/n$이다.

④ O
기대도수는 행의 표본크기 n_i에 공통 범주비율 $O_{\cdot j}/n$을 곱해 구한다. 따라서 i번째 모집단의 j번째 범주 추정 기대도수는 $n_i O_{\cdot j}/n$이다.

오답 포인트

①은 합을 취하는 방향을 바꿔 놓은 선지이다. 동질성검정에서는 p_{ij}가 '모집단 i 내부의 범주비율'이라는 뜻이므로 1이 되는 합은 열방향이 아니라 같은 모집단 내부 범주합이다. 행과 열의 의미를 뒤바꾸면 바로 오답이 된다.

다시 나오면 이렇게 판단

먼저 p_{ij}가 무엇의 비율인지부터 정한다. 그다음 귀무가설에서는 '모집단마다 같은 범주비율'인지 확인하고, 공통 비율은 $O_{\cdot j}/n$, 기대도수는 '행합 × 열비율'인 $n_i O_{\cdot j}/n$으로 연결하면 된다.

암기 포인트

동질성검정은 '분포가 같은가'를 본다. 한 모집단 안에서는 $\sum_{j} p_{ij} = 1$, 공통 비율은 $O_{\cdot j}/n$, 기대도수는 $n_i O_{\cdot j}/n$이다.

NoteBase

문 7. 질병에 대한 치료효과를 검정하기 위해 환자 10명의 치료 전과 치료 후의 검사치가 다음 표와 같다. 여기에서 $X_1, X_2, \cdots, X_{10}$은 서로 독립이며 정규분포 $N(\mu_X, \sigma_X^2)$를, $Y_1, Y_2, \cdots, Y_{10}$은 서로 독립이며 정규분포 $N(\mu_Y, \sigma_Y^2)$를 따른다. 치료 전후 검사치의 차이의 모평균 μ_D에 대해 귀무가설 $H_0 : \mu_D = 0$ 대 대립가설 $H_1 : \mu_D > 0$을 유의수준 α에서 검정하고자 한다. 이에 대한 설명으로 옳은 것은? (단, $t_\alpha(n)$는 자유도 n인 t 분포의 제 $100 \times (1-\alpha)$ 백분위수이다)

시기 \ 환자	1	2	3	⋯	10	평균	분산
치료 전	X_1	X_2	X_3	⋯	X_{10}	$\overline{X}$	S_X^2
치료 후	Y_1	Y_2	Y_3	⋯	Y_{10}	$\overline{Y}$	S_Y^2
차이($D_i = X_i - Y_i$)	D_1	D_2	D_3	⋯	D_{10}	$\overline{D}$	S_D^2

① 검정통계량 $\frac{\overline{D}}{S_D/\sqrt{10}}$이 $-t_\alpha(9)$보다 작으면 귀무가설을 기각한다.
② 검정통계량 $\frac{\overline{D}}{S_D/\sqrt{10}}$이 $t_\alpha(9)$보다 크면 귀무가설을 기각한다.
③ 검정통계량 $\frac{\overline{D}}{S_D/\sqrt{10}}$이 $-t_\alpha(10)$보다 작으면 귀무가설을 기각한다.
④ 검정통계량 $\frac{\overline{D}}{S_D/\sqrt{10}}$이 $t_\alpha(10)$보다 크면 귀무가설을 기각한다.

정답 2번

출제 포인트

같은 환자의 치료 전후 차이 $D_i = X_i - Y_i$를 이용하는 대응표본 t검정에서 검정통계량, 자유도, 우측단측 기각역을 정확히 잡는지가 핵심이다.

조건 정리

주어진 값: 환자 10명에 대해 $D_i = X_i - Y_i$, 표본평균은 $\overline{D}$, 표본표준편차는 S_D이다.
구해야 할 것: $H_0 : \mu_D = 0$ 대 $H_1 : \mu_D > 0$의 기각역이다.
주의할 조건: 같은 환자의 전후 차이를 쓰므로 대응표본 검정이고, 표본 수가 10이므로 자유도는 $10 - 1 = 9$이다. 대립가설이 $\mu_D > 0$이므로 우측단측 검정이다.

NoteBase

사용할 공식

검정통계량은 $T = \frac{\overline{D}}{S_D/\sqrt{10}}$이다.
귀무가설이 참이면 $T \sim t(9)$로 본다.
따라서 유의수준 α의 기각역은 $T > t_\alpha(9)$이다.
이 공식을 쓰는 이유는 두 평균을 따로 비교하는 문제가 아니라 같은 10명의 전후 차이의 평균 μ_D가 0보다 큰지를 묻는 문제이기 때문이다.

계산 과정

1. 환자별 차이 $D_i = X_i - Y_i$를 정의하므로 대응표본 t검정을 사용한다.
2. 표본 수가 10이므로 자유도는 $10 - 1 = 9$이다.
3. 대립가설이 $H_1 : \mu_D > 0$이므로 오른쪽 꼬리에서 기각한다.
4. 따라서 $\frac{\overline{D}}{S_D/\sqrt{10}} > t_\alpha(9)$이면 귀무가설을 기각한다.

선지별 해설

① X: $\frac{\overline{D}}{S_D/\sqrt{10}} < -t_\alpha(9)$는 좌측단측 기각역이다. 이 문제는 $H_1 : \mu_D > 0$이므로 틀린 표현은 "$-t_\alpha(9)$보다 작으면"이고, 올바르게는 "$t_\alpha(9)$보다 크면"이다.
② O: 대응표본 t검정의 통계량은 $\frac{\overline{D}}{S_D/\sqrt{10}}$, 자유도는 9, 대립가설은 우측단측이므로 기각역은 $T > t_\alpha(9)$이다.
③ X: 방향도 틀리고 자유도도 틀렸다. 틀린 표현은 "$-t_\alpha(10)$보다 작으면"이고, 올바르게는 "$t_\alpha(9)$보다 크면"이다.
④ X: 우측단측이라는 방향은 맞지만 자유도를 10으로 둔 점이 틀렸다. 틀린 표현은 "$t_\alpha(10)$"이고, 올바르게는 "$t_\alpha(9)$"이다.

오답 포인트

대응표본 검정인데도 독립 2표본처럼 생각하거나, 자유도를 표본 수 10으로 그대로 쓰는 실수가 잦다. 또 대립가설의 부호를 놓치면 우측단측 검정을 좌측단측 검정으로 뒤집어 ①처럼 판단하게 된다.

다시 나오면 이렇게 판단

전후 비교 문항은 먼저 차이변수 D_i를 정의했는지 확인하고, 그다음 대립가설의 부호로 꼬리 방향을 정한 뒤, 대응표본 t검정의 자유도 $n - 1$를 넣어 기각역을 쓰면 된다.

암기 포인트

대응표본 t검정은 차이의 평균을 본다. 검정통계량은 $\frac{\overline{D}}{S_D/\sqrt{n}}$, 자유도는 $n - 1$, 기각 방향은 대립가설의 부호가 결정한다.

문 8. 다음은 어떤 자료를 다중선형회귀모형 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$에 적합하여 얻은 결과의 일부이다. 이에 대한 설명으로 옳지 않은 것은? (단, 오차항 e는 평균이 0이고 분산이 σ^2인 정규분포를 따른다)

모형	비표준화 계수	표준화 계수	⋯	유의확률
상수	1.793		⋯	
X_1	1.971	0.816	⋯	0.000
X_2	−1.290	−1.337	⋯	0.000

① Y에 X_2가 X_1보다 더 크게 영향을 미친다.

② X_1, X_2의 회귀계수는 유의수준 5%에서 모두 유의하다.

③ X_1이 고정되어 있을 때 X_2가 증가하면 Y는 감소하는 경향이 있다.

④ X_2가 고정되어 있을 때 X_1이 1단위 증가하면 Y의 추정값이 0.816배 증가한다.

정답 4번

출제 포인트

이 문제는 다중선형회귀표에서 비표준화계수, 표준화계수, 유의확률의 의미를 구분하는 문제이다.

자료 읽는 순서

1. 비표준화계수에서 각 설명변수가 1단위 변할 때 Y 추정값이 얼마나 변하는지 본다.
2. 표준화계수의 절댓값으로 변수 간 상대적 영향력 크기를 비교한다.
3. 유의확률이 0.05보다 작은지 확인해 각 회귀계수의 유의성을 판단한다.

NoteBase

선지별 해설

① O

표준화계수는 단위 차이를 제거한 영향력 비교 지표이다. X_1의 표준화계수는 0.816, X_2의 표준화계수는 -1.337이므로 절댓값 기준으로 X_2의 영향력이 더 크다.

② O

유의확률이 두 변수 모두 0.000이므로 유의수준 5%에서 X_1, X_2의 회귀계수는 모두 유의하다.

③ O

비표준화계수에서 X_2의 계수는 -1.290이다. 따라서 X_1을 고정하면 X_2가 1단위 증가할 때 Y의 추정값은 1.290만큼 감소한다.

④ X

0.816은 X_1의 표준화계수이므로 1단위 증가 효과를 뜻하지 않는다. X_2를 고정할 때 X_1이 1단위 증가하면 Y의 추정값은 비표준화계수 1.971만큼 증가한다. 틀린 표현은 '0.816배 증가한다'이고, 올바르게는 'X_1이 1단위 증가할 때 Y의 추정값이 1.971만큼 증가한다'이다.

오답 포인트

표준화계수와 비표준화계수를 섞어 읽으면 ④처럼 단위 변화 효과를 잘못 해석하게 된다. 표준화계수는 변수 간 상대적 영향력 비교용이고, 1단위 증가에 따른 Y의 변화량은 비표준화계수로 읽는다.

다시 나오면 이렇게 판단

회귀표 문제에서는 먼저 유의확률로 유의성을 확인하고, 다음으로 비표준화계수로 1단위 변화 효과의 방향과 크기를 읽는다. 변수 간 영향력 비교를 묻는 경우에만 표준화계수의 절댓값을 사용한다.

암기 포인트

1단위 변화 효과는 비표준화계수, 영향력 비교는 표준화계수이다.

NoteBase

문 9. $X_1, \cdots, X_n$**은 확률밀도함수** $f(x;\mu)=\frac{1}{\mu}e^{-\frac{x}{\mu}}$, $(\mu>0, x>0)$**을 따르는 확률표본이다. 모수** μ**의 최대가능도추정량(maximum likelihood estimator)** $\hat{\mu}$**에 대한 설명으로 옳지 않은 것은? (단, log는 자연로그이다)**

① 가능도함수(likelihood function)를 대수변환한 로그가능도함수는 $l(\mu)=-n\log\mu-\frac{1}{\mu}\sum_{i=1}^{n}x_i$이다.
② 로그가능도함수를 μ에 대해 한 번 미분해서 0으로 놓고, 이 방정식의 해를 구한다.
③ 최대가능도추정량은 $\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}X_i$이다.
④ 최대가능도추정량은 편의추정량(biased estimator)이다.

정답 4번

출제 포인트

지수분포 $f(x;\mu)=\frac{1}{\mu}e^{-x/\mu}$에서 로그가능도함수를 만들고, 이를 미분해 최대가능도추정량 $\hat{\mu}$를 구한 뒤 그 추정량의 불편성을 판단하는 문제이다.

조건 정리

주어진 분포는 평균이 μ인 지수분포이다.
표본은 $X_1, \ldots, X_n$이다.
발문은 '옳지 않은 것'이므로 틀린 설명 하나를 찾는 문제이다.

사용할 공식

가능도함수는
$L(\mu)=\prod_{i=1}^{n}\frac{1}{\mu}e^{-x_i/\mu}=\mu^{-n}e^{-\sum_{i=1}^{n}x_i/\mu}$
이고, 로그가능도함수는
$l(\mu)=\log L(\mu)=-n\log\mu-\frac{1}{\mu}\sum_{i=1}^{n}x_i$
이다.
이 문항은 최대가능도추정량을 묻고 있으므로 로그가능도함수를 μ에 대해 미분한 score equation을 사용한다. 또한 구해진 추정량의 편의 여부는 기댓값 $E[\hat{\mu}]$를 확인해 판단한다.

NoteBase

계산 과정

1. $l(\mu) = -n\log\mu - \frac{1}{\mu}\sum_{i=1}^{n} x_i$를 구한다.
2. 이를 μ에 대해 미분하면
$\frac{dl}{d\mu} = -\frac{n}{\mu} + \frac{1}{\mu^2}\sum_{i=1}^{n} x_i$
이다.
3. $\frac{dl}{d\mu} = 0$으로 두면
$-n\mu + \sum_{i=1}^{n} x_i = 0$
이므로
$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$
이다.
4. 이차미분은
$\frac{d^2l}{d\mu^2} = \frac{n}{\mu^2} - \frac{2}{\mu^3}\sum_{i=1}^{n} x_i$
이고, $\mu = \bar{X}$에서 음수가 되어 최대점이 된다.
5. 따라서 최대가능도추정량은 $\bar{X}$이고, $E[\hat{\mu}] = E[\bar{X}] = \mu$이므로 불편추정량이다.

선지별 해설

① O
로그를 취하면 곱이 합으로 바뀌므로 $l(\mu) = -n\log\mu - \frac{1}{\mu}\sum x_i$가 된다. 제시된 로그가능도함수 식은 정확하다.

② O
최대가능도추정량의 후보는 로그가능도함수를 μ에 대해 미분해 0으로 두는 score equation으로 구한다. 이 문항에서는 그 해가 $\hat{\mu} = \bar{X}$이고, 이차미분값이 음수이므로 실제 최대점이 된다.

③ O
②의 방정식을 풀면 $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$가 된다. 이 선지는 최대가능도추정량의 값을 바르게 쓴 것이다.

④ X
틀린 표현은 '편의추정량'이다. 이 문항의 최대가능도추정량은 $\hat{\mu} = \bar{X}$이고, 지수분포의 평균이 μ이므로 $E[\bar{X}] = \mu$이다. 따라서 $\hat{\mu}$는 편의추정량이 아니라 불편추정량이다. 올바르게는 '최대가능도추정량은 불편추정량이다.'가 된다.

오답 포인트

지수분포의 모수화를 혼동하면 쉽게 틀린다. $f(x;\mu) = \frac{1}{\mu}e^{-x/\mu}$에서는 μ가 rate가 아니라 평균이므로 최대가능도추정량이 $1/\bar{X}$가 아니라 $\bar{X}$가 된다. 또한 'MLE는 항상 편의가 있다'고 일반화하면 ④에 걸리게 된다.

NoteBase

다시 나오면 이렇게 판단

지수분포 문제는 먼저 모수가 평균형 μ인지 rate형 λ인지 확인한다. 그다음 로그가능도함수를 만든 뒤 score equation으로 $\hat{\mu}$를 구하고, 마지막에 $E[\hat{\mu}]$를 계산해 편의 여부를 확인하면 된다. 발문이 '옳은 것'인지 '옳지 않은 것'인지도 함께 확인해야 한다.

암기 포인트

지수분포 $f(x;\mu)=\frac{1}{\mu}e^{-x/\mu}$에서는 평균이 μ이고, 최대가능도추정량은 $\bar{X}$, 이 문항에서는 불편추정량이다.

문 10. X_1, X_2, X_3, X_4, X_5**는 서로 독립이며, 구간** $(0, 1)$**에서 균일분포 (uniform distribution)를 따르고,** $X_{(1)} \le X_{(2)} \le X_{(3)} \le X_{(4)} \le X_{(5)}$**는** X_1, X_2, X_3, X_4, X_5**의 순서통계량이라고 할 때, 기대값** $E[X_{(4)}]$**는?**

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

정답 3번

출제 포인트

이 문제는 균일분포의 순서통계량 기대값 공식 $E[X_{(k)}] = \frac{k}{n+1}$을 이용해 $E[X_{(4)}]$를 구하는 문제이다.

조건 정리

- 주어진 값: X_1, X_2, X_3, X_4, X_5는 서로 독립이고 $(0, 1)$에서 균일분포를 따른다.
- 순서통계량: $X_{(1)} \le X_{(2)} \le X_{(3)} \le X_{(4)} \le X_{(5)}$이다.
- 구해야 할 값: $E[X_{(4)}]$이다.
- 주의할 조건: 표본 수는 5개이므로 분모는 $5 + 1$이 된다.

사용할 공식

균일분포 $(0, 1)$의 k번째 순서통계량 기대값은 $E[X_{(k)}] = \frac{k}{n+1}$이다.

이 공식을 쓰는 이유:
이 문제는 독립인 균일분포 표본의 순서통계량 기대값을 직접 묻고 있으므로, 순서통계량의 표준 결과를 그대로 적용할 수 있다.

계산 과정

1. 표본 수는 $n = 5$, 구하려는 순서는 $k = 4$로 잡는다.
2. 공식에 대입하면 $E[X_{(4)}] = \frac{4}{5+1}$이다.
3. 따라서 $E[X_{(4)}] = \frac{4}{6} = \frac{2}{3}$이다.

NoteBase

선지별 해설

① X

$\frac{1}{6}$은 $\frac{1}{n+1}$ 값으로, 첫째 순서통계량 $E[X_{(1)}]$에 해당한다. 이 문제는 네 번째 순서통계량을 묻고 있으므로 틀리다. 틀린 표현은 값의 순서 선택이다. 올바르게는 $E[X_{(4)}] = \frac{4}{6}$이다.

② X

$\frac{1}{3} = \frac{2}{6}$은 둘째 순서통계량 $E[X_{(2)}]$에 해당한다. $k = 4$를 넣어야 하는데 $k = 2$로 잘못 넣은 값이므로 틀리다. 틀린 표현은 분자의 순서 번호이다. 올바르게는 분자가 4가 되어 $\frac{4}{6} = \frac{2}{3}$이다.

③ O

$E[X_{(4)}] = \frac{4}{6} = \frac{2}{3}$이므로 맞다. 표본 5개 중 네 번째로 작은 값의 기대위치는 0과 1 사이를 6등분했을 때 네 번째 점에 해당한다.

④ X

$\frac{5}{6}$은 다섯째 순서통계량 $E[X_{(5)}]$의 기대값이다. 최대값의 기대값과 네 번째 순서통계량의 기대값을 혼동한 선택지이므로 틀리다. 틀린 표현은 순서통계량 번호이다. 올바르게는 최대값이 아니라 네 번째 값이므로 $\frac{4}{6} = \frac{2}{3}$이다.

오답 포인트

이 문항의 함정은 순서통계량 번호 k와 표본 수 n을 섞는 데 있다. ①은 가장 작은 값, ②는 두 번째 값, ④는 가장 큰 값의 기대값이라서 모두 익숙한 숫자이지만, 발문은 정확히 $X_{(4)}$를 묻는다.

다시 나오면 이렇게 판단

균일분포 순서통계량 문제는 먼저 표본 수 n과 순서 번호 k를 확인한 뒤, $E[X_{(k)}] = \frac{k}{n+1}$에 그대로 대입해 판단한다. 오답 선지는 다른 순서통계량 $E[X_{(1)}]$, $E[X_{(2)}]$, $E[X_{(5)}]$ 값인지 함께 점검하면 된다.

암기 포인트

균일분포 순서통계량의 기대값은 $\frac{k}{n+1}$이다. 분자는 순서 번호, 분모는 표본 수에 1을 더한 값이다.

NoteBase

문 11. 전체 인구의 5%가 어떤 질병을 가지고 있다고 한다. 이 질병을 진단하는 검사 방법이 있는데, 이 검사 방법은 이 질병을 실제로 가지고 있을 때 검사 결과가 양성(positive) 반응을 나타낼 확률이 0.95, 음성(negative) 반응을 나타낼 확률이 0.05이고, 이 질병을 실제로 가지고 있지 않을 경우 검사 결과가 양성 반응을 나타낼 확률이 0.1, 음성 반응을 나타낼 확률이 0.9이다. 임의로 한 사람을 선택하여 이 검사를 한 결과가 양성 반응으로 나타났다면, 이 사람이 실제로 이 질병을 가지고 있을 확률은?

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

정답 2번

출제 포인트

이 문제는 베이즈 정리를 이용해 양성 판정을 받은 사람이 실제로 질병을 가질 조건부확률을 구하는 문제이다.

조건 정리

- 유병률: 0.05
- 민감도 $P(+|D)$: 0.95
- 위양성률 $P(+|D^c)$: 0.1
- 구해야 할 값: $P(D|+)$

사용할 공식

$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$
이 공식을 쓰는 이유는 문제에서 양성 반응이 나왔다는 조건 아래 실제 질병 보유 확률을 묻고 있기 때문이다.

계산 과정

1. 질병이 있을 사전확률은 $P(D) = 0.05$, 없을 확률은 $P(D^c) = 0.95$이다.
2. 실제 환자가 양성일 확률은 $P(+|D)P(D) = 0.95 \times 0.05 = 0.0475$이다.
3. 비환자가 양성일 확률은 $P(+|D^c)P(D^c) = 0.1 \times 0.95 = 0.095$이다.
4. 따라서 전체 양성 확률은 $P(+) = 0.0475 + 0.095 = 0.1425$이다.
5. 그러므로
$P(D|+) = \frac{0.0475}{0.1425} = \frac{1}{3}$
이 된다.

NoteBase

선지별 해설

① X: $\frac{1}{2}$는 정답이 아니다. 양성 판정자 중에는 위양성도 포함되므로 실제 질병 확률은 절반보다 작다. 틀린 표현은 값 자체이며, 올바르게는 $\frac{1}{3}$이다.
② O: 계산 결과 $P(D \mid +) = \frac{1}{3}$이므로 옳다.
③ X: $\frac{1}{4}$는 정답이 아니다. 이는 양성자 중 실제 환자 비중을 과소평가한 값이다. 틀린 표현은 값 자체이며, 올바르게는 $\frac{1}{3}$이다.
④ X: $\frac{1}{5}$는 정답이 아니다. 유병률 5%만 보고 사후확률도 낮게 잡으면 이런 오답이 나오기 쉽다. 틀린 표현은 값 자체이며, 올바르게는 $\frac{1}{3}$이다.

오답 포인트

유병률 $P(D)$와 양성 후 확률 $P(D \mid +)$를 같은 것으로 착각하면 오답이 된다. 이 문항의 함정은 검사 정확도가 높아 보여도 위양성률 0.1과 비환자 집단 95%가 함께 작용하면 양성자 중 실제 환자 비율이 크게 낮아진다는 점이다.

다시 나오면 이렇게 판단

검사 문제에서 '양성 반응이 나왔을 때 실제로 질병이 있을 확률'을 묻는다면 사후확률 문제로 보고, 민감도와 위양성률에 각각 사전확률을 곱한 뒤 양성 전체 확률로 나누어 판단한다.

암기 포인트

베이즈 정리는 '관심 사건의 양성 몫 ÷ 전체 양성'으로 정리된다.

NoteBase

문 12. 여론조사 회사에서 어떤 사항에 대한 국민의 찬성 비율 p를 알아보기 위해 전화 설문조사를 무작위로 실시하려고 한다. 96% 신뢰수준에서 모든 p에 대한 추정량의 오차한계를 0.05이내로 하는 최소 표본의 크기 n을 구하는 부등식은? (단, z_α는 표준정규분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $z_{0.04} = 1.75$, $z_{0.02} = 2.05$이고, 모집단의 크기는 충분히 크다)

① $n \geq \left(\frac{1.75}{0.05}\right)^2 \frac{21}{2}$
② $n \geq \left(\frac{1.75}{0.05}\right)^2 \frac{21}{4}$
③ $n \geq \left(\frac{2.05}{0.05}\right)^2 \frac{21}{2}$
④ $n \geq \left(\frac{2.05}{0.05}\right)^2 \frac{21}{4}$

정답 4번

출제 포인트

비율 p의 신뢰구간 오차한계 공식을 쓰고, 모든 p에서 오차한계를 보장하려면 $p(1-p)$의 최댓값 1/4를 써야 하는 문제이다.

조건 정리

- 신뢰수준은 96%이다.
- 오차한계는 0.05 이내이어야 한다.
- 모든 p에 대해 성립해야 한다.
- $z_{0.04} = 1.75$, $z_{0.02} = 2.05$이다.

사용할 공식

비율 추정의 오차한계는 $E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ 이다.
96% 신뢰수준이면 $\alpha = 0.04$ 이므로 $\alpha/2 = 0.02$, 따라서 $z_{\alpha/2} = z_{0.02} = 2.05$ 이다.
또한 모든 p에 대해 오차한계를 보장하려면 $p(1-p)$의 최댓값 1/4를 넣어야 한다.

계산 과정

$0.05 \geq z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
$0.05 \geq 2.05\sqrt{\frac{1/4}{n}}$
양변을 제곱하면
$0.05^2 \geq 2.05^2 \frac{1}{4n}$
정리하면
$n \geq \left(\frac{2.05}{0.05}\right)^2 \frac{21}{4}$
따라서 정답은 4번이다.

NoteBase

선지별 해설

① X: 1.75를 쓴 부분이 틀리다. 96% 신뢰수준에서는 $\alpha = 0.04$ 이고 표준정규분포 분위수는 $z_{\alpha/2} = z_{0.02} = 2.05$ 를 써야 한다. 올바르게는 $n \geq \left(\frac{2.05}{0.05}\right)^2 \frac{1}{4}$ 이다.
② X: 1/4를 쓴 것은 맞지만 분위수를 1.75로 둔 것이 틀리다. 올바르게는 $z_{0.02} = 2.05$ 를 써서 $n \geq \left(\frac{2.05}{0.05}\right)^2 \frac{1}{4}$ 이다.
③ X: 분위수 2.05 는 맞지만 $p(1-p)$의 최댓값을 1/2로 본 것이 틀리다. $p(1-p)$의 최댓값은 $p = 1/2$에서 1/4 이다. 올바르게는 $n \geq \left(\frac{2.05}{0.05}\right)^2 \frac{1}{4}$ 이다.
④ O: 96% 신뢰수준이므로 $z_{\alpha/2} = z_{0.02} = 2.05$ 를 쓰고, 모든 p에 대해 보장하려면 $p(1-p) \leq 1/4$ 를 이용하므로 $n \geq \left(\frac{2.05}{0.05}\right)^2 \frac{1}{4}$ 가 된다.

오답 포인트

96% 신뢰수준에서 바로 $z_{0.04}$를 쓰는 실수가 자주 나온다. 신뢰구간의 양쪽 꼬리를 나누므로 항상 $\alpha/2$를 본다. 또 모든 p에 대해 보장한다는 조건을 놓치면 $p(1-p)$의 최댓값 1/4 대신 다른 값을 넣는 실수가 나온다.

다시 나오면 이렇게 판단

비율의 표본크기 문제는 먼저 신뢰수준을 α와 $\alpha/2$로 바꿔 임계값을 정하고, 다음으로 모든 p 조건이 있으면 $p(1-p) = 1/4$를 넣은 뒤, 마지막에 오차한계 식을 n에 대해 정리하면 된다.

암기 포인트

모든 비율 p에 대한 표본크기는 $n \geq \frac{z_{\alpha/2}^2}{4E^2}$ 꼴로 정리된다.

NoteBase

문 13. 어느 가게에 월요일부터 금요일까지 방문하는 손님 수는 다음 표와 같다. 손님 수가 요일에 따라 다른지를 검정하기 위한 χ^2 검정통계량의 값과 유의수준 5%에서 검정결과를 바르게 연결한 것은? (단, $\chi^2_\alpha(k)$는 자유도가 k인 χ^2 분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $\chi^2_{0.05}(4) = 9.49$, $\chi^2_{0.05}(5) = 11.07$이다)

요일	월	화	수	목	금	합계
손님 수(명)	25	25	10	15	25	100

① χ^2 검정통계량의 값: 10, 검정결과: 귀무가설을 기각할 수 없음

② χ^2 검정통계량의 값: 10, 검정결과: 귀무가설을 기각함

③ χ^2 검정통계량의 값: 11, 검정결과: 귀무가설을 기각할 수 없음

④ χ^2 검정통계량의 값: 11, 검정결과: 귀무가설을 기각함

정답 2번

출제 포인트

이 문제는 카이제곱 적합도 검정에서 기대도수를 구한 뒤 검정통계량과 임계값을 비교해 귀무가설 기각 여부를 판단하는 문제이다.

조건 정리

관측도수는 월 25, 화 25, 수 10, 목 15, 금 25이다.
귀무가설 H_0: 요일에 따라 손님 수가 다르지 않다.
전체 100명을 5개 요일에 균등하게 나누면 기대도수는 각 요일 20명이다.
자유도는 범주 수 5에서 1을 뺀 4이다.

사용할 공식

검정통계량은 $\chi^2 = \sum \frac{(O-E)^2}{E}$ 이다.
이 공식을 쓰는 이유는 요일별 관측도수와 귀무가설 아래의 기대도수가 얼마나 차이 나는지를 카이제곱 값으로 측정하는 적합도 검정 상황이기 때문이다.

NoteBase

계산 과정

각 요일의 기대도수는 $100/5 = 20$이다.

따라서

$$\chi^2 = \frac{(25-20)^2}{20} + \frac{(25-20)^2}{20} + \frac{(10-20)^2}{20} + \frac{(15-20)^2}{20} + \frac{(25-20)^2}{20}$$

$$= \frac{25}{20} + \frac{25}{20} + \frac{100}{20} + \frac{25}{20} + \frac{25}{20} = 10$$

자유도 4에서 유의수준 5%의 임계값은 $\chi^2_{0.05}(4) = 9.49$이다.

검정통계량 10은 9.49보다 크므로 귀무가설을 기각한다.

1번

X. 검정통계량의 값 10은 맞지만, "귀무가설을 기각할 수 없음"이 틀렸다. 실제로는 $10 > 9.49$이므로 귀무가설을 기각한다. 틀린 표현은 "귀무가설을 기각할 수 없음"이고, 올바르게는 "귀무가설을 기각함"이다.

2번

O. 검정통계량은 10이고, 자유도 4의 임계값 9.49보다 크므로 귀무가설을 기각한다. 검정통계량의 값과 검정결과가 모두 올바르게 연결되어 있다.

3번

X. "검정통계량의 값 11"이 틀렸고 검정결과도 틀렸다. 실제 검정통계량은 10이며, $10 > 9.49$이므로 귀무가설을 기각한다. 틀린 표현은 "11"과 "귀무가설을 기각할 수 없음"이고, 올바르게는 "10"과 "귀무가설을 기각함"이다.

4번

X. "귀무가설을 기각함"이라는 결론 자체는 맞지만, 검정통계량을 11로 쓴 부분이 틀려서 선택지 전체는 틀리다. 실제 값은 10이다. 틀린 표현은 "11"이고, 올바르게는 "10"이다.

오답 포인트

이 문항의 함정은 검정통계량 계산과 기각역 판단을 따로 보지 않고 한 번에 찍는 데 있다. 1번은 통계량 10은 맞지만 결론을 뒤집은 선택지이고, 4번은 결론은 맞지만 통계량 자체를 잘못 쓴 선택지이다. 계산값과 임계값 비교를 분리해서 확인해야 걸리지 않는다.

다시 나오면 이렇게 판단

적합도 검정 문제는 먼저 귀무가설 아래 기대도수를 구하고, 다음에 $\chi^2 = \sum (O-E)^2/E$로 검정통계량을 계산한 뒤, 마지막에 자유도와 임계값을 비교해 기각 여부를 판단하면 된다. 계산값이 맞아도 결론이 틀린 선택지가 자주 섞이므로 둘을 따로 점검해야 한다.

NoteBase

암기 포인트

적합도 검정은 기대도수부터 구하고, χ^2값이 임계값보다 크면 귀무가설을 기각한다.

NoteBase

문 14. 다음은 처리 수준의 수가 5인 공정에서 처리에 따른 반응값의 모평균이 차이가 있는지를 검정하기 위한 일원배치 분산분석법을 적용한 결과의 일부이다. 검정에 대한 유의확률(significance probability) p의 범위는? (단, $F_\alpha(k_1, k_2)$는 분자의 자유도가 k_1이고 분모의 자유도가 k_2인 F분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $F_{0.01}(4, 17) = 4.67$, $F_{0.05}(4, 17) = 2.96$, $F_{0.1}(4, 17) = 2.31$이다)

요인	제곱합	자유도	평균제곱	F-값
처리	220	4		
잔차				
전체	390	21		

① $0 < p < 0.01$

② $0.01 < p < 0.05$

③ $0.05 < p < 0.1$

④ $0.1 < p < 1$

정답 1번

출제 포인트

이 문제는 일원배치 분산분석표에서 평균제곱과 F값을 계산한 뒤, 그 F값을 임계값과 비교하여 유의확률 p의 범위를 판단하는 문제이다.

조건 정리

주어진 값은 처리 제곱합 220, 처리 자유도 4, 전체 제곱합 390, 전체 자유도 21이다.
잔차 제곱합은 $390 - 220 = 170$이고 잔차 자유도는 $21 - 4 = 17$이다.
구해야 할 값은 검정통계량 F와 그에 대응하는 유의확률 p의 범위이다.

사용할 공식

일원배치 분산분석에서는 처리 평균제곱 $MS_T = SS_T/df_T$, 잔차 평균제곱 $MS_E = SS_E/df_E$, 검정통계량 $F = MS_T/MS_E$를 사용한다.
이 공식을 쓰는 이유는 문제에서 처리 수준이 5인 공정의 모평균 차이를 검정하는 일원배치 분산분석 결과표를 주었기 때문이다.

NoteBase

계산 과정

1. 잔차 제곱합과 잔차 자유도를 구하면 $SS_E = 390 - 220 = 170$, $df_E = 21 - 4 = 17$이다.
2. 처리 평균제곱은 $MS_T = 220/4 = 55$이다.
3. 잔차 평균제곱은 $MS_E = 170/17 = 10$이다.
4. 따라서 검정통계량은 $F = 55/10 = 5.5$이다.
5. $F = 5.5$를 임계값과 비교하면 $5.5 > F_{0.01}(4, 17) = 4.67$이므로 상측확률인 유의확률 p는 0.01보다 작다.

선지별 해설

① O: 계산된 F값이 5.5이고 이는 $F_{0.01}(4, 17) = 4.67$보다 크므로 $p < 0.01$이다. 따라서 ①의 범위는 옳다.
② X: 틀린 표현은 "0.01 < p < 0.05"이다. 실제로는 F값이 1% 임계값까지 이미 넘었으므로 올바르게는 "0 < p < 0.01"이다.
③ X: 틀린 표현은 "0.05 < p < 0.1"이다. 이 범위가 되려면 F값이 $F_{0.1}(4, 17) = 2.31$과 $F_{0.05}(4, 17) = 2.96$ 사이에 있어야 하는데, 실제 $F = 5.5$는 훨씬 크다. 올바르게는 "0 < p < 0.01"이다.
④ X: 틀린 표현은 "0.1 < p < 1"이다. 이 범위는 검정통계량이 매우 작아 기각 근거가 약할 때의 경우인데, 실제 $F = 5.5$는 제시된 모든 임계값보다 크다. 올바르게는 "0 < p < 0.01"이다.

오답 포인트

전체 제곱합에서 처리 제곱합을 빼서 잔차 제곱합을 구하지 않으면 평균제곱 계산이 무너진다. 또 분산분석의 유의확률은 상측확률이므로 F값이 클수록 p가 작아진다는 방향을 반대로 해석하면 ②, ③, ④ 같은 오답을 고르게 된다.

다시 나오면 이렇게 판단

분산분석표 문제는 먼저 빈칸인 잔차 제곱합과 자유도를 채운 다음, 평균제곱 두 개를 구하고, 마지막에 F값을 임계값과 크기 비교하여 p의 구간을 판단하는 순서로 정리하면 된다.

암기 포인트

일원배치 분산분석은 F=처리 평균제곱/잔차 평균제곱이고, F가 클수록 유의확률 p는 작아진다.

NoteBase

문15. 관측된 자료 $(x_i, y_i), i = 1, \cdots, 21$에서 얻은 통계량의 값이 $\sum_{i=1}^{21}(x_i - \bar{x})^2 = 81, \quad \sum_{i=1}^{21}(y_i - \bar{y})^2 = 100$이다. 이 자료를 단순선형 회귀모형 $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$에 적합하여 얻은 기울기에 대한 최소제곱추정값이 $\hat{\beta}_1 = -1$일 때, 이에 대한 설명으로 옳은 것은? (단, 오차항 ε_i는 서로 독립이며 정규분포 $N(0, \sigma^2)$을 따른다)

① X가 증가할 때 Y가 증가하는 경향이 있다.

② 설명변수와 반응변수의 상관계수는 0.9이다.

③ 회귀모형의 결정계수는 0.81이다.

④ 오차항의 분산 σ^2에 대한 추정값인 평균제곱오차는 10이다.

정답 3번

출제 포인트

이 문제는 단순회귀에서 기울기 추정치로부터 상관계수, 결정계수, 평균제곱오차를 연결해 판단할 수 있는지를 묻는 문제이다.

조건 정리

주어진 값은 $S_{xx} = \sum_{i=1}^{21}(x_i - \bar{x})^2 = 81, S_{yy} = \sum_{i=1}^{21}(y_i - \bar{y})^2 = 100, \hat{\beta}_1 = -1$이다.
표본크기는 21이므로 오차자유도는 $21 - 2 = 19$이다.
구해야 할 것은 각 선지의 진위이다.

사용할 공식

$\hat{\beta}_1 = S_{xy}/S_{xx}, r = S_{xy}/\sqrt{S_{xx}S_{yy}}, R^2 = r^2, SSR = \hat{\beta}_1 S_{xy}, SSE = S_{yy} - SSR, MSE = SSE/(n-2)$를 사용한다.
기울기 추정치와 제곱합이 함께 주어졌으므로 먼저 S_{xy}를 구하면 나머지 통계량이 연쇄적으로 정해진다.

계산 과정

1. $\hat{\beta}_1 = S_{xy}/S_{xx} = -1$이므로 $S_{xy} = -81$이다.
2. 상관계수는 $r = -81/\sqrt{81 \cdot 100} = -81/90 = -0.9$이다.
3. 결정계수는 $R^2 = r^2 = (-0.9)^2 = 0.81$이다.
4. 회귀제곱합은 $SSR = \hat{\beta}_1 S_{xy} = (-1)(-81) = 81$, 오차제곱합은 $SSE = 100 - 81 = 19$이다.
5. 평균제곱오차는 $MSE = 19/19 = 1$이다.

NoteBase

선지 판단

1번: X. 기울기 추정치가 -1이므로 X가 증가할수록 Y는 감소하는 경향이 있다. 틀린 표현은 "증가"이며, 올바르게는 "감소"이다.
2번: X. 상관계수는 $r = -0.9$이다. 틀린 표현은 "0.9"이며, 올바르게는 "-0.9"이다.
3번: O. 단순회귀에서 결정계수는 $R^2 = r^2$이므로 0.81이 맞다.
4번: X. 평균제곱오차는 $MSE = 1$이다. 틀린 표현은 "10"이며, 올바르게는 "1"이다.

오답 포인트

기울기와 상관계수의 부호를 절댓값과 분리해서 보지 않으면 ②를 맞다고 오인하기 쉽다. 또 $S_{yy} = 100$을 오차제곱합으로 착각하면 ④처럼 평균제곱오차를 잘못 계산하게 된다.

다시 나오면 이렇게 판단

단순회귀 수치문제는 먼저 $\hat{\beta}_1$와 S_{xx}로 S_{xy}를 복원한 뒤, 부호가 필요한 통계량은 r, 제곱이 필요한 통계량은 R^2, 오차 관련 통계량은 $SSR \to SSE \to MSE$ 순서로 판단하면 된다.

암기 포인트

단순회귀에서는 $\hat{\beta}_1 = S_{xy}/S_{xx}$, $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$, $R^2 = r^2$, $MSE = SSE/(n-2)$의 연결고리를 함께 기억해 두면 된다.

NoteBase

문 16. 처리 수준의 수가 4인 인자 A와 처리 수준의 수가 2인 인자 B에 대한 반복이 없는 이원배치법의 실험에서 다음과 같은 자료를 얻었다. 분산분석 결과에서 인자 A와 인자 B가 모두 유의수준 5%에서 유의한 것으로 판명되었다. A_2수준과 B_2수준을 조합한 실험 조건에서 모평균의 추정값은?

인자 A / 인자 B	A_1	A_2	A_3	A_4	합계
B_1	80	38	98	56	272
B_2	42	22	70	42	176
합계	122	60	168	98	448

① 18

② 22

③ 30

④ 44

정답 1번

출제 포인트

반복이 없는 이원배치법에서 교호작용을 두지 않는 가법모형으로 특정 셀의 모평균을 추정하는 문제이다.

조건 정리

- 인자 A는 4수준, 인자 B는 2수준이다.
- 반복이 없으므로 각 조합의 관측치는 1개뿐이다.
- 구해야 할 값은 A_2와 B_2를 조합한 조건의 모평균 추정값이다.

사용할 공식

반복이 없는 이원배치법의 적합값은

$\hat{\mu}_{ij} = \bar{y}_{i\cdot} + \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$

이다.

이 공식을 쓰는 이유:

반복이 없으면 교호작용을 따로 추정할 수 없으므로 가법모형에서 해당 A수준 평균과 해당 B수준 평균을 더한 뒤 전체평균을 한 번 빼서 셀 평균을 추정한다.

NoteBase

계산 과정

1. A_2의 평균은 $60/2 = 30$이다.
2. B_2의 평균은 $176/4 = 44$이다.
3. 전체평균은 $448/8 = 56$이다.
4. 따라서 A_2B_2 조건의 모평균 추정값은 $30 + 44 - 56 = 18$이다.

선지별 해설

① O

- A_2 평균 30, B_2 평균 44, 전체평균 56을 이용하면 추정값은 $30 + 44 - 56 = 18$이다.

② X

- 22는 A_2B_2 셀에서 실제로 관측된 값이다. 문제는 관측값이 아니라 가법모형에 따른 모평균 추정값을 묻는다.
- 틀린 표현은 '추정값이 22이다'이다. 올바르게는 '관측값은 22이지만 모평균 추정값은 18이다'이다.

③ X

- 30은 A_2 열평균만 계산한 값이다. B_2 효과와 전체평균 보정이 빠져 있으므로 셀 평균 추정값이 아니다.
- 틀린 표현은 '추정값이 30이다'이다. 올바르게는 'A_2 수준 평균은 30이고, A_2B_2 모평균 추정값은 18이다'이다.

④ X

- 44는 B_2 행평균만 계산한 값이다. 특정 셀의 추정값이 아니라 B_2 수준 전체의 평균이다.
- 틀린 표현은 '추정값이 44이다'이다. 올바르게는 'B_2 수준 평균은 44이고, A_2B_2 모평균 추정값은 18이다'이다.

오답 포인트

이 문항의 함정은 셀 관측값 22를 그대로 답으로 고르거나, A_2 평균 30 또는 B_2 평균 44만 보고 멈추는 데 있다. 반복이 없는 이원배치법에서 특정 조합의 추정평균은 관측값 하나로 정하지 않고 두 주효과와 전체평균 보정을 함께 반영해야 한다.

다시 나오면 이렇게 판단

반복이 없는 이원배치법에서 특정 조합의 추정평균을 묻으면 먼저 해당 A수준 평균, 해당 B수준 평균, 전체평균을 구한 뒤 해당 A수준 평균+해당 B수준 평균-전체평균 순서로 계산한다.

암기 포인트

반복 없는 이원배치의 셀 추정값은 해당 두 수준의 평균을 더하고 전체평균을 한 번 빼는 값이다.

NoteBase

문 17. 숫자 1, 2, 3, 4가 표시된 공정한 정사면체 A, B 두 개를 던져서 나오는 눈의 수를 각각 a, b라고 하고, 확률변수 $X = \left\lfloor \frac{b}{a} \right\rfloor$라고 할 때, 확률변수 $Y = 16X$의 기댓값은? (단, $[x]$는 x보다 크지 않은 최대 정수를 나타낸다)

① 15

② 16

③ 17

④ 18

정답 3번

출제 포인트

이 문제는 $X = \left\lfloor \frac{b}{a} \right\rfloor$의 기대값을 경우별로 구한 뒤, 기대값의 선형성으로 $Y = 16X$의 기대값을 계산하는 문제이다.

조건 정리

주어진 값: a, b는 각각 1, 2, 3, 4를 같은 확률로 갖는다.
구해야 할 값: $E(Y)$
주의할 조건: $[x]$는 x 이하의 최대 정수이므로 X는 $\frac{b}{a}$ 자체가 아니라 $\left\lfloor \frac{b}{a} \right\rfloor$이다.

사용할 공식

기대값의 선형성에 따라 $E(Y) = E(16X) = 16E(X)$이다.
또한 a와 b의 16가지 경우가 모두 동일한 확률이므로, $X = \left\lfloor \frac{b}{a} \right\rfloor$의 값을 모두 구해 평균을 내면 된다.
이 공식을 쓰는 이유: 문제는 먼저 X의 평균값을 구한 다음 상수 16을 곱하는 구조이기 때문이다.

계산 과정

1. $a = 1$이면 $X = [b]$이므로 1,2,3,4가 되어 평균은 $\frac{1+2+3+4}{4} = \frac{5}{2}$이다.
2. $a = 2$이면 $X = \left\lfloor \frac{b}{2} \right\rfloor$이므로 0,1,1,2가 되어 평균은 $\frac{4}{4} = 1$이다.
3. $a = 3$이면 $X = \left\lfloor \frac{b}{3} \right\rfloor$이므로 0,0,1,1이 되어 평균은 $\frac{2}{4} = \frac{1}{2}$이다.
4. $a = 4$이면 $X = \left\lfloor \frac{b}{4} \right\rfloor$이므로 0,0,0,1이 되어 평균은 $\frac{1}{4}$이다.
5. 따라서 $E(X) = \frac{1}{4}\left(\frac{5}{2} + 1 + \frac{1}{2} + \frac{1}{4}\right) = \frac{17}{16}$이다.
6. 그러므로 $E(Y) = 16E(X) = 16 \cdot \frac{17}{16} = 17$이다.

NoteBase

선지별 해설

① X: 계산 결과 $E(Y) = 17$이므로 15는 틀리다. 틀린 표현은 '15'이고, 올바르게는 '17'이다. 이는 16가지 경우의 X 값을 덜 합산한 경우에 나오는 오답이다.
② X: 계산 결과 $E(Y) = 17$이므로 16은 틀리다. 틀린 표현은 '16'이고, 올바르게는 '17'이다. 이는 $\left\lfloor \frac{b}{a} \right\rfloor$ 값을 단순화하여 $E(X) = 1$로 잘못 본 경우에 나오는 값이다.
③ O: 계산 결과 $E(X) = \frac{17}{16}$, $E(Y) = 17$이므로 맞다. 바닥함수를 각 경우에 정확히 적용하면 최종값은 17이다.
④ X: 계산 결과 $E(Y) = 17$이므로 18은 틀리다. 틀린 표현은 '18'이고, 올바르게는 '17'이다. 이는 $\left\lfloor \frac{4}{3} \right\rfloor$ 같은 값을 과대평가하거나 일부 경우를 잘못 더한 경우에 나오는 오답이다.

오답 포인트

출제자가 노린 함정은 $[x]$를 일반 나눗셈값으로 보는 실수이다. 예를 들어 $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}$ 같은 값은 그대로 더하는 것이 아니라 모두 바닥함수를 취해 0으로 처리해야 한다. 또 $E(16X)$를 구할 때 먼저 X의 평균을 정확히 구한 뒤 16을 곱해야 한다.

다시 나오면 이렇게 판단

바닥함수 기대값 문제는 먼저 각 경우의 값을 정수로 바꾼 뒤 평균을 구하는 순서로 판단한다. 특히 분모 역할을 하는 값별로 묶어 $\left\lfloor \frac{b}{a} \right\rfloor$ 표를 만들면 계산 누락을 막을 수 있다.

암기 포인트

기대값은 상수배가 밖으로 나오고, 바닥함수는 평균보다 먼저 적용된다는 점이 핵심이다.

NoteBase

문 18. $X_1, X_2, \cdots, X_{10}$**은 평균이 0, 분산이 4인 정규모집단으로부터 추출한 확률표본이고,** $Y_1, Y_2, \cdots, Y_{20}$**은 평균이 0, 분산이 8인 정규모집단으로부터 추출한 확률표본이다. 두 모집단이 독립일 때, 표본평균** $\overline{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$**와** $\overline{Y} = \frac{1}{20}\sum_{i=1}^{20} Y_i$**를 이용한 통계량** $\left(\frac{\overline{X}+\overline{Y}}{\overline{X}-\overline{Y}}\right)^2$**의 확률분포는?**

① 자유도가 1인 카이제곱분포

② 자유도가 10인 카이제곱분포

③ 분자의 자유도가 1이고 분모의 자유도가 1인 F분포

④ 분자의 자유도가 10이고 분모의 자유도가 20인 F분포

정답 3번

출제 포인트

독립인 두 정규표본의 표본평균을 선형결합한 뒤, 공분산이 0인 공동정규의 독립성과 F분포의 정의를 연결하는 문제이다.

조건 정리

$\bar{X} \sim N(0, 4/10) = N(0, 2/5)$이고 $\bar{Y} \sim N(0, 8/20) = N(0, 2/5)$이다.
두 표본은 독립이므로 $\bar{X}$와 $\bar{Y}$도 독립이다.
구해야 할 통계량은 $\left(\frac{\bar{X}+\bar{Y}}{\bar{X}-\bar{Y}}\right)^2$의 분포이다.

사용할 공식

정규확률변수의 선형결합은 다시 정규분포를 따르고, 공동정규에서 공분산이 0이면 서로 독립이다.
또한 서로 독립인 $Z_1, Z_2 \sim N(0,1)$에 대해 $\frac{Z_1^2/1}{Z_2^2/1} \sim F(1,1)$이다.
이 문제는 분자와 분모를 각각 표준정규로 바꾼 뒤 제곱비 꼴로 볼 수 있으므로 이 공식을 사용한다.

계산 과정

1. $U = \bar{X} + \bar{Y}, V = \bar{X} - \bar{Y}$로 놓으면 U, V는 정규분포를 따른다.
2. $\mathrm{Var}(U) = \mathrm{Var}(\bar{X}) + \mathrm{Var}(\bar{Y}) = 2/5 + 2/5 = 4/5$이고, $\mathrm{Var}(V) = 2/5 + 2/5 = 4/5$이다.
3. $\mathrm{Cov}(U, V) = \mathrm{Cov}(\bar{X}+\bar{Y}, \bar{X}-\bar{Y}) = \mathrm{Var}(\bar{X}) - \mathrm{Var}(\bar{Y}) = 2/5 - 2/5 = 0$이므로 U와 V는 서로 독립이다.
4. 따라서 $Z_1 = U/\sqrt{4/5}$, $Z_2 = V/\sqrt{4/5}$는 서로 독립인 표준정규분포를 따른다.
5. 그러므로
$\left(\frac{\bar{X}+\bar{Y}}{\bar{X}-\bar{Y}}\right)^2 = \left(\frac{U}{V}\right)^2 = \frac{Z_1^2}{Z_2^2} = \frac{Z_1^2/1}{Z_2^2/1} \sim F(1,1)$
이다.

NoteBase

선지별 해설

① X: 자유도 1인 카이제곱분포는 표준정규 한 개를 제곱한 Z^2 꼴이다. 이 문제의 통계량은 그런 제곱 한 개가 아니라 두 독립 카이제곱의 비이므로 카이제곱분포가 아니다. 틀린 표현은 "카이제곱분포"이고, 올바르게는 "자유도 1과 1인 F분포"이다.
② X: 자유도 10은 원표본의 크기에서 바로 가져온 값인데, 여기서는 표본평균의 선형결합을 표준화한 뒤 제곱비를 보는 문제이므로 자유도가 10이 되지 않는다. 틀린 표현은 "자유도가 10인 카이제곱분포"이고, 올바르게는 "표준화한 $\bar{X}+\bar{Y}$ 또는 $\bar{X}-\bar{Y}$ 각각의 제곱은 자유도 1인 카이제곱분포"이다.
③ O: $\bar{X}+\bar{Y}$와 $\bar{X}-\bar{Y}$는 서로 독립이고, 둘 다 분산이 4/5인 정규분포이다. 따라서 각각을 $\sqrt{4/5}$로 나누면 독립인 표준정규가 되고, 그 제곱의 비는 $F(1,\,1)$분포를 따른다.
④ X: $F(10,\,20)$은 보통 두 독립 표본분산을 나눈 형태에서 나오는 자유도이다. 이 문제는 표본분산이 아니라 표본평균의 선형결합 두 개를 비교하므로 자유도가 10과 20으로 가지 않는다. 틀린 표현은 "분자의 자유도가 10이고 분모의 자유도가 20인 F분포"이고, 올바르게는 "분자와 분모의 자유도가 각각 1인 F분포"이다.

오답 포인트

표본크기 10과 20을 그대로 자유도로 옮기는 실수가 가장 흔하다. 이 문항에서 자유도는 원자료 개수가 아니라 최종적으로 만들어진 독립 표준정규 변수의 개수로 정해진다. 또 비를 제곱했다고 해서 바로 카이제곱분포로 오해하기 쉬운데, 제곱 하나는 카이제곱이고 제곱 둘의 비는 F분포이다.

다시 나오면 이렇게 판단

평균의 선형결합 비율이 나오면 먼저 각 결합이 정규인지 확인하고, 분산과 공분산을 계산해 독립 여부를 본다. 그다음 각각을 표준화해서 표준정규의 제곱 한 개인지, 두 제곱의 비인지 구분하면 χ^2, t, F 중 어느 분포인지 바로 정리된다.

암기 포인트

독립인 표준정규 두 개의 제곱비는 $F(1,\,1)$이다. 표본크기는 평균의 비 문제에서 자유도로 바로 들어가지 않는다.

NoteBase

문 19. 어느 자료에서 변수 X와 Y를 평균이 0, 표준편차가 1이 되도록 각각 표준화한 변수를 X^*와 Y^*라 하자. 설명변수 X^*와 반응변수 Y^*를 절편이 있는 단순선형회귀모형에 적용하여 얻은 추정회귀식과 분산분석표에 대한 설명으로 옳은 것만을 모두 고르면?

ㄱ. 추정회귀식의 절편은 0이다.
ㄴ. 분산분석표에서 회귀제곱합과 오차제곱합의 합은 회귀자유도와 오차자유도의 합과 같다.
ㄷ. X와 Y가 각각 설명변수와 반응변수인 절편이 있는 단순선형회귀모형에 적용하여 얻은 분산분석표의 F-값과 X^*와 Y^*를 이용한 분산분석표의 F-값은 같다.

① ㄱ
② ㄱ, ㄴ
③ ㄴ, ㄷ
④ ㄱ, ㄴ, ㄷ

정답 4번

출제 포인트

표준화된 설명변수와 반응변수로 단순선형회귀를 할 때 절편, 제곱합과 자유도의 관계, 그리고 표준화 전후 F값의 불변성을 함께 묻는 문제이다.

핵심 개념

표준화한 변수 X^*, Y^*는 평균이 0이고 표준편차가 1이다.
절편이 있는 단순선형회귀에서 절편은 $\bar{Y}^* - b\bar{X}^*$이므로 두 평균이 0이면 절편도 0이다.
또한 Y^*의 총제곱합은 표본표준편차 기준으로 $n-1$이고, 회귀제곱합과 오차제곱합의 합은 총제곱합과 같으므로 자유도 합 $1+(n-2)=n-1$와 대응된다.
마지막으로 F값은 상관계수 또는 결정계수 R^2로 결정되는데 표준화는 상관구조를 바꾸지 않으므로 F값도 변하지 않는다.

NoteBase

선지별 해설

ㄱ은 O이다. 추정회귀식의 절편은 $\hat{\beta}_0 = \bar{Y}^* - \hat{\beta}_1 \bar{X}^*$인데 $\bar{X}^* = 0,\ \bar{Y}^* = 0$이므로 절편은 0이다.
ㄴ은 O이다. 분산분석표에서 회귀제곱합과 오차제곱합의 합은 총제곱합이고, Y^*의 표본표준편차가 1이면 총제곱합은 $n-1$이다. 한편 회귀자유도는 1, 오차자유도는 $n-2$이므로 그 합도 $n-1$이다.
ㄷ은 O이다. 원자료 X, Y로 회귀하든 표준화한 X^*, Y^*로 회귀하든 상관계수와 결정계수 R^2는 같고, 단순선형회귀의 F값은 $F = \frac{R^2}{1-R^2}(n-2)$로 표현되므로 동일하다.
① X이다. ①은 ㄱ만 옳다고 본 선택지인데 실제로는 ㄱ, ㄴ, ㄷ이 모두 옳다. 틀린 표현은 'ㄱ만'에 해당하는 조합이다. 올바르게는 ④처럼 ㄱ, ㄴ, ㄷ을 모두 포함해야 한다.
② X이다. ②는 ㄷ을 빠뜨린 조합이다. ㄷ도 표준화 전후 F값이 같다는 점에서 옳으므로 올바르게는 ④가 된다.
③ X이다. ③은 ㄱ을 빠뜨린 조합이다. ㄱ의 절편 0 설명도 옳으므로 올바르게는 ④가 된다.
④ O이다. ㄱ, ㄴ, ㄷ이 모두 옳으므로 옳은 것만 모두 고른 선택지는 ④이다.

오답 포인트

ㄴ에서 '제곱합의 합이 자유도의 합과 같다'는 문장을 무조건 성립하는 일반명제로 읽으면 흔들릴 수 있다. 이 문항은 Y^*를 표준화하여 총제곱합이 $n-1$이 되도록 만든 특수한 상황을 전제로 한다는 점이 함정이다. 또 ㄷ은 회귀계수 자체는 바뀌더라도 검정통계량 F는 변하지 않는다는 점을 놓치기 쉽다.

다시 나오면 이렇게 판단

표준화 회귀 문제는 먼저 평균 0 여부로 절편을 본다. 다음으로 표준화된 반응변수의 총제곱합이 얼마인지 확인한다. 마지막으로 표준화가 상관계수와 R^2를 바꾸지 않는다는 점을 이용해 t값과 F값의 불변성을 판단한다.

암기 포인트

표준화 회귀는 절편 0, 총제곱합 $n-1$, R^2와 F는 불변이다.

NoteBase

문 20. 다음은 어느 제품에 포함된 두 인자 A, B의 수준에 따라 제품의 품질이 어떤 영향을 받는지 알아보기 위해 반복이 있는 이원배치 모형을 적용하여 얻은 분산분석표의 일부이다. 재분석 결과에 대한 설명으로 옳지 않은 것은? (단, $F_\alpha(k_1, k_2)$는 분자의 자유도가 k_1이고 분모의 자유도가 k_2인 F분포의 제$100 \times (1-\alpha)$ 백분위수를 나타내고, $F_{0.05}(1, 4) = 7.71$, $F_{0.05}(1, 5) = 6.61$이다)

요인	제곱합	자유도	평균제곱	$F-$값
인자 A	16	1		
인자 B	14	1		
교호작용 $A \times B$	2			
오차	8			
전체		7		

여기서 교호작용의 효과가 유의수준 5%에서 유의하지 않아 교호작용을 포함하지 않는 이원배치 모형을 적용하여 재분석을 실시하였다. 다음은 재분석하여 얻은 분산분석표의 일부이다.

요인	제곱합	자유도	평균제곱	$F-$값
인자 A	16	1		
인자 B	14	1		
오차				
전체				

① 유의수준 5%에서 인자 A의 효과가 있다고 할 수 있다.

② 유의수준 5%에서 인자 B의 효과가 있다고 할 수 있다.

③ 오차항의 자유도는 4이다.

④ 평균제곱오차는 2이다.

정답 3번

NoteBase

출제 포인트

이 문제는 반복이 있는 이원배치 분산분석에서 유의하지 않은 교호작용을 오차항에 합쳐 재분석할 때 자유도, 평균제곱오차, 검정통계량이 어떻게 바뀌는지를 묻는다.

조건 정리

처음 분산분석표에서 인자 A의 제곱합은 16, 인자 B의 제곱합은 14, 교호작용의 제곱합은 2, 오차의 제곱합은 8이다.
전체 자유도는 7이고, 인자 A와 B의 자유도는 각각 1이므로 교호작용 자유도는 $(1)(1)=1$이다.
따라서 처음 오차 자유도는 $7-1-1-1=4$이다.
교호작용이 유의하지 않으므로 재분석에서는 교호작용 제곱합과 자유도를 오차항에 합친다.

사용할 공식

평균제곱은 $MS=SS/df$, 검정통계량은 $F=MS_{\text{factor}}/MS_E$이다.
교호작용을 제외한 뒤에는 $SS_E'=SS_E+SS_{AB}$, $df_E'=df_E+df_{AB}$를 사용한다.
이 공식을 쓰는 이유는 유의하지 않은 교호작용을 별도 효과로 보지 않고 오차에 합쳐 주효과를 다시 검정하기 때문이다.

계산 과정

처음 분석에서 오차 자유도는 4, 평균제곱오차는 $MS_E=8/4=2$이다.
교호작용의 자유도는 1이므로 재분석 오차 자유도는 $df_E'=4+1=5$이다.
재분석 오차제곱합은 $SS_E'=8+2=10$이므로 평균제곱오차는 $MS_E'=10/5=2$이다.
따라서 재분석에서 인자 A의 F값은 $16/2=8$, 인자 B의 F값은 $14/2=7$이다.
주어진 임계값과 비교하면 $8>F_{0.05}(1,5)=6.61$, $7>F_{0.05}(1,5)=6.61$이므로 인자 A와 인자 B는 모두 유의하다.

선지별 해설

1번: O. 재분석에서 평균제곱오차가 2이므로 인자 A의 F값은 $16/2=8$이다. $8>6.61$이므로 유의수준 5%에서 인자 A의 효과가 있다고 할 수 있다.
2번: O. 재분석에서 인자 B의 F값은 $14/2=7$이다. $7>6.61$이므로 유의수준 5%에서 인자 B의 효과가 있다고 할 수 있다.
3번: X. 재분석에서는 교호작용 자유도 1을 오차항에 합치므로 오차항 자유도는 4가 아니라 5이다. 틀린 표현은 '오차항의 자유도는 4이다'이고, 올바르게는 '오차항의 자유도는 5이다'이다.
4번: O. 재분석 오차제곱합은 $8+2=10$, 오차 자유도는 5이므로 평균제곱오차는 $10/5=2$이다. 따라서 이 선지는 옳다.

NoteBase

오답 포인트

처음 분석의 오차 자유도 4를 재분석에도 그대로 쓰면 3번을 옳다고 착각하기 쉽다. 교호작용을 제거하는 것이 교호작용 항을 버리는 것이 아니라 그 제곱합과 자유도를 오차항에 합치는 절차라는 점이 핵심 함정이다.

다시 나오면 이렇게 판단

이원배치 분산분석에서 유의하지 않은 교호작용을 제외하면 먼저 교호작용의 제곱합과 자유도를 오차항에 더한 뒤 새 평균제곱오차를 구하고, 그 값으로 주효과의 F값을 다시 계산해 임계값과 비교한다.

암기 포인트

유의하지 않은 교호작용은 버리는 항이 아니라 오차에 합치는 항이므로 재분석에서는 SS_E와 df_E가 함께 증가한다.

NoteBase