수리통계학 교재 목차

제 1 장　확률 이론

Def (1): 표본공간(Sample space)

통계적 실험에서 발생가능한 모든 결과들의 집합을 표본공간이라하고 이를 S로 나타낸다.

예제 1

다음 각 경우의 표본공간을 구하여라.

(1) 한 개의 동전을 던져서 앞면이 나오면 다시 동전을 던지고 뒷면이 나오면 한 개의 주사위를 던지는 실험 (동전 앞면 H, 뒷면 T로 표기하기로 한다.)

(2) 한 개의 동전을 3번 던지는 실험

(3) 한 개의 동전을 3번 던질 때 앞면이 나오는 횟수의 실험

(4) 비누공장의 생산 제품 중에서 세 개의 제품을 추출하여 검사한 뒤 불량품과 정품을 분류하는 경우 (여기서 불량품을 F, 제품은 H로 표시하기로 한다.)

(5) 중심이 원점이고 반지름이 3인 원의 원주상 및 내부의 모든 점

Def (2) : 사상(events)

사상은 표본공간의 부분집합이다.

예제 2

비누 완제품 중에서 3개를 추출하여 검사한 후 정품과 불량품으로 구별할 때 불량품의 개수가 1개 이상인 사상은?

Thm (1) : 사상들과의 관계

(1) 전사상(total event)

⇒

(2) 공사상(null event)

⇒

(3) 여사상(complementary event)

⇒

(4) 합사상(union event)

⇒

(5) 곱사상(intersection event)

⇒

(6) 배반사상(mutually exclusive event)

⇒

예제 3

한 개의 주사위는 실험에서 $E=\{1,\ 2,\ 3\}$, $F=\{1,\ 3,\ 5\}$라 할 때 표본공간, 두 사상의 합사상, 곱사상, E의 여사상, 총사상의 개수를 구하여라.

예제 4

한 개의 동전을 던져서 앞면이 나오는 면 H, 뒷면이 나오는 면 T라 할 때

(1) 표본공간 S를 구하여라.

(2) 첫 번째 던진 동전이 앞면이 나오는 사상 E를 구하여라.

(3) 두 번째 던진 동전이 뒷면이 나오는 사상 F를 구하여라.

(4) E와 F의 합사상, 곱사상을 구하여라.

(5) E와 F는 배반인지 판단하여라.

Thm (2) : 곱의 법칙

원소의 개수가 n_1, n_2, $\cdots$, n_k인 집합 A_1, A_2, $\cdots$, A_k에서 각각 하나의 원소를 택하여 나열한 순서열의 개수는 $n_1 n_2 n_3 \cdots n_k$이다.

예제 5

10개의 O, X 문제를 한 학생이 풀 때 대답가능한 가지수는?

예제 6

한 개의 동전과 두 개의 주사위를 던질 때 나오는 경우의 수는?

예제 7

0, 1, 2, 3, 4, 5를 각각 한 번씩 사용하여 만들 수 있는 4자리 짝수의 개수는?

예제 8

차의 번호판이 두 개의 다른 알파벳 문자와 첫 번째 숫자가 0이 아닌 세 개의 숫자가 차례로 구성된다면 몇 개의 번호판이 가능한가?

※ 수형도를 이용한 카운팅

예제 9

A팀과 B팀이 야구게임을 할 때 어느 팀이 연속적 2게임을 이기거나 전부 4게임을 먼저 이기는 팀이 우승한다고 한다. 발생가능한 방법수는?

예제 10

한 사람이 최대 5번을 할 수 있는 게임이 있다. 각 게임에서 이기면 1달러를 받고 지면 1달러를 잃는다. 이 사람은 1달러를 가지고 게임을 시작하고 돈을 모두 잃거나 3달러를 따면 (즉, 4달러가 되면) 5경기 전에도 멈춘다. 이 사람이 5경기 전에 멈추는 경우의 수는?

Thm (3) : 순열(Permutation)

서로 다른 n개의 원소 중에서 r개를 뽑아 나열하는 방법의 수는

$${}_nP_r = n(n-1)\times \cdots \times(n-r+1) = \frac{n!}{(n-r)!} \text{ (비복원 추출)}$$

만약 중복이 허락되면 ${}_n\Pi_r = n^r$ (복원 추출)

※ 동자순열 ⇒

※ 원순열 ⇒

예제 11

a, b, c, d, e 중에서 세 개를 선택하여 만들 수 있는 단어의 수는?

예제 12

10장의 카드에서 3장을 뽑아 나열하는 방법수를 구하여라.

(1) 비복원 추출

(2) 복원 추출

예제 13

4개의 붉은 색 깃발과 2개의 청색 깃발을 모두 사용하여 만들어지는 서로 다른 신호의 개수는?

Thm (4) : 조합(combination)

서로 다른 n개에서 r개를 선택하는 방법수는

$$ {}_nC_r = \binom{n}{r} = \frac{{}_nP_r}{r!} = \frac{n!}{r!(n-r)!} \quad \text{(동시추출)} $$

예제 14

어떤 사람이 필기구를 구입하려고 한다. A사 제품 10개 중에서 3개, B사 제품 5개 중에서 2개를 선택하는 방법수는?

예제 15

$\binom{n+1}{r} = \binom{n}{r-1} + \binom{n}{r}$ 임을 증명하여라.

Def (3) : 확률의 공리적 정의

사상 A의 확률은 사상 A안에 있는 모든 표본점에 할당된 확률의 합이다. 따라서, $0 \le P(A) \le 1$, $P(\phi) = 0$, $P(S) = 1$이다.
또, A_1, A_2, A_3, $\cdots$ 가 상호배반인 사상이면

$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots = \sum_{i=1}^{\infty} P(A_i)$$ 이다.

예제 16

짝수가 홀수보다 3배만큼 더 많이 발생하는 주사위를 한 번 던질 때 4보다 작은 수가 나올 사상을 A라 할 때 $P(A)$를 구하여라.

예제 17

짝수가 홀수보다 2배만큼 더 많이 발생하는 주사위를 한 번 던질 때 짝수가 나타날 사상을 A, 3의 배수가 나타날 사상을 B라 할 때 $P(A \cup B)$와 $P(A \cap B)$를 구하여라.

Thm (5)

어떤 시행에서 N개의 서로 다른 결과가 동일한 확률로 일어날 때
이 중 n개의 원소를 가지는 사상 E의 확률은

$P(E)=\dfrac{n}{N}$ 이다.

예제 18

주사위를 두 개 던질 때 그 눈이 합이 4이하일 확률을 구하여라.

예제 19

흰 공 6개, 검은 공 4개가 들어있는 상자에서 공 3개를 임의로 꺼낼 때 검은 공이 두 개 포함될 확률은?

예제 20

3개의 볼트와 3개의 너트가 들어있는 상자에서 2개를 임의로 꺼낼 때 볼트 1개, 너트 1개일 확률은?

Thm (6) : 확률의 기본정리

(1) $P(E^c) = 1 - P(E)$

(2) E와 F가 서로 배반이 아니면

$P(E \cup F) = P(E) + P(F) - P(E \cap F)$

$\Rightarrow$

예제 21

$P(A)=0.7$, $P(B)=0.5$, $P(A\cap B)=0.3$ 일 때

(1) $P(A^c)=$

(2) $P(B^c)=$

(3) $P(A^c\cap B)=$

(4) $P(A\cup B)=$

(5) $P(A\cap B^c)=$

(6) $P(A\cup B^c)=$

(7) $P(A^c\cap B^c)=$

(8) $P(A^c\cup B^c)=$

예제 22

$P(A\cup B)=\frac{7}{8}$, $P(A\cap B)=\frac{1}{4}$, $P(A^c)=\frac{5}{8}$ 일 때 $P(A\cap B^c)$ 를 구하여라.

예제 23

15 개의 부품 중에 5 개의 불량품이 들어 있다.

임의로 3 개를 꺼낼 때 적어도 하나의 불량품이 포함될 확률은?

Thm (7) : 가법정리(additive rule)의 확장

$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{k=1}^{n} (-1)^{k-1} S_k$ 이다.

여기서 $S_1 = P(A_1) + P(A_2) + \cdots + P(A_n) \Rightarrow \binom{n}{1}$개

$S_2 = P(A_1 \cap A_2) + \cdots + P(A_{n-1} \cap A_n) \Rightarrow \binom{n}{2}$개

$S_3 = P(A_1 \cap A_2 \cap A_3) + \cdots + P(A_{n-2} \cap A_{n-1} \cap A_n) \Rightarrow \binom{n}{3}$개

$\vdots$

$S_n = P(A_1 \cap A_2 \cap \cdots \cap A_n) \Rightarrow \binom{n}{n}$개

① $P(A_1 \cup A_2) =$

② $P(A_1 \cup A_2 \cup A_3) =$

③ $P(A_1 \cup A_2 \cup A_3 \cup A_4) =$

④ $P(A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5) =$

예제 24

카드는 총 52장으로 4가지의 무늬로 구성되어 있다. 52장의 카드에서 임의로 13장을 뽑을 때 같은 무늬의 에이스와 킹을 뽑을 확률을 구하여라.

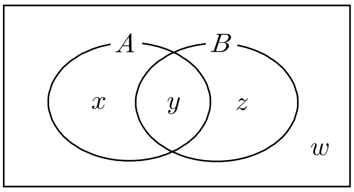
예제 25

5명의 수험생에게 임의로 수험표를 나누어줄 때 적어도 한 명은 자기 수험표를 받게 될 확률을 구하여라.

Def (4): 조건부 확률

한 사상이 일어났다는 전제 하에서 다른 사상이 일어날 확률을 조건부 확률이라 한다.
사상 A가 발생했다는 전제하에서 사상 B가 일어날 조건부 확률은
$P(B/A)=\dfrac{P(A\cap B)}{P(A)}$ 로 정의한다. (단, $P(A)>0$)

$\Rightarrow$



예제 26
주사위를 두 개 던져서 두 눈의 합이 10이 나오는 경우를 사상 E라 하고 서로 다른 눈이 나오는 경우를 사상 F라 할 때 $P(F/E)$를 구하여라.

예제 27
한 쌍의 주사위를 던져 눈의 합이 6일 때 적어도 한 번 2의 눈이 나올 확률을 구하여라.

예제 28

흰 공 3개와 검은 공 2개가 들어있는 주머니에서 하나씩 차례로 두 개의 공을 꺼낼 때 첫 번째 꺼낸 공이 검은 공이었을 때 두 번째 꺼낸 공도 검은 공일 확률을 구하여라.

예제 29

어느 비누 공장에서 비누를 제조했을 때 무게 미달 불량 10%, 크기불량 5%, 두 부분 모두 불량일 확률이 0.8%라 할 때 이 공장에서 생산된 비누를 임의로 하나 선택하여 검사한 결과, 무게 불량이었을 때 이것이 크기 불량일 확률은?

Def (5) : 독립과 종속

$P(B/A) = P(B)$ 이거나 $P(A/B) = P(A)$ 이면 두 사상 A 와 B 는 독립이고 그렇지 않으면 종속이라 한다.

Thm (8) : 승법정리(multiplicative rule)

$P(A \cap B) = P(A) \cdot P(B/A)$ 이다.

만약, A 와 B 가 독립이면 $P(A \cap B) = P(A) \cdot P(B)$ 이다.

예제 30

52장의 카드에서 카드를 한 장씩 두 번 선택할 때 다음 물음에 답하여라.

(1) 복원 추출했을 때 2장 모두 스페이스일 확률은?

(2) 비복원 추출했을 때 2장 모두 스페이스일 확률은?

예제 31

빨간 공 2개, 흰 공 3개가 들어 있는 주머니에서 한 개의 공을 꺼내어 그 색깔을 확인 후 같은 색깔의 공을 하나 더 주머니에 넣은 다음 두 번째 공을 꺼낼 때 꺼낸 공이 모두 빨간 공일 확률은?

예제 32

목표물을 명중할 확률이 각각 $\frac{1}{2}$, $\frac{2}{3}$ 인 두 사람이 있다. 두 사람이 목표물에 사격할 때 명중할 확률은?

예제 33

임의의 두 사상 A, B에 대하여 A와 B가 독립이면 A와 B^c, A^c와 B, A^c와 B^c도 각각 독립임을 보여라.

예제 34

A와 B가 독립이고 $P(A)=0.6$, $P(B)=0.3$일 때 확률 $P(A\cup B^c)$를 구하여라.

예제 35

A, B, C가 서로 독립이면 $A \cap B$와 C^c, A와 $B \cup C$, A와 $B \cap C$도 서로 독립임을 증명하여라.

예제 36

1에서 30까지의 자연수 중에서 하나를 뽑을 때 사상 $A=\{$짝수$\}$,
사상 $B=\{3$의 배수$\}$, 사상 $C=\{5$의 배수$\}$일 때 A, B, C는 서로 독립인가?

Thm (9)

어떤 실험에서 사상 A_1, A_2, $\cdots$, A_k가 발생가능하다면

$$P(A_1 \cap A_2 \cap \cdots \cap A_k)$$
$$= P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2) \times \cdots \times P(A_k/A_1 \cap A_2 \cap \cdots \cap A_{k-1})$$

이 성립한다.

※ 만약 사상 A_1, A_2, $\cdots$, A_k가 서로 독립이면

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1)P(A_2) \times \cdots \times P(A_k)$$

$\Rightarrow$

예제 37

52장의 카드에서 한 장씩 차례로 4장 뽑을 때 2, 5, 7, 7이 나올 확률을 구하여라.

예제 38

검은 공 9개, 흰 공 1개가 들어있는 주머니에서 비복원추출로 공을 계속해서 1개씩 꺼낼 때 4번째 꺼낸 공이 흰 공일 확률을 구하여라.

Thm (10): 전체확률법칙

사상 $B_1,\ B_2,\ \cdots,\ B_k$를 표본공간 S의 분할이라 하고

$P(B_i) \neq 0,\ i=1,\ 2,\ \cdots,\ k$라면 S의 임의의 사상 A에 대하여

$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i) P(A/B_i)$가 성립한다.

$\Rightarrow$

예제 39

세 개의 생산라인에서 각각 전체 생산품의 50%, 30%, 20%를 만들고 각각 2%, 5%, 10%의 불량품이 생긴다. 임의로 하나의 생산품을 선택했을 때 50%의 생산라인일 사상을 A, 30% 생산라인일 사상을 B, 20% 생산라인일 사상을 C라 하자. 이 때, 임의로 선택한 제품이 불량품일 사상을 D라 할 때 $P(D)$를 구하여라.

예제 40

빨간 공 2개, 흰 공 5개가 들어있는 상자와 빨간공 3개, 흰 공 4개가 들어있는 상자가 있다. 첫 번째 상자에서 하나의 공을 꺼내어 두 번째 상자에 넣고 두 번째 상자에서 하나의 공을 꺼낼 때 빨간공이 나올 확률은?

Thm (11): 베이즈 정리(Bay's Theorem)

B_1, B_2, $\cdots$, B_n 이 표본공간 S의 분할이고 A가 S의 임의의 사상이라면

$$P(B_i/A)=\frac{P(B_i)\cdot P(A/B_i)}{\sum_{i=1}^{n}P(B_i)\cdot P(A/B_i)} \quad (\text{단, } P(B_i)>0,\ P(A)>0)$$

$\Rightarrow$

예제 41

어떤 공장은 세 개의 생산라인 B_1, B_2, B_3에서 각각 전체 제품의 30%, 45%, 25%를 만들고 또한 각각 2%, 3%, 2%의 불량품이 생긴다.

이 때, 임의로 선택한 하나의 생산품이 불량품이었을 때 그 제품이 생산라인 B_3에서 만든 제품일 확률을 구하여라.

예제 42

상자 속에 3장의 카드가 있다. 하나는 양면이 모두 빨간색이고 두 번째는 양면이 모두 파란색이며 세 번째는 한 면은 빨간색, 다른 면은 파란색이다. 상자에서 1장의 빨간색 카드를 꺼낼 때 그 뒷면도 빨간색일 확률은?

예제 43

어떤 병원의 암 검사의 적중률은 95%이다. 즉, 실제 암환자를 검사결과 양성으로 판단할 확률은 0.95이다. 또 암에 걸리지 않은 사람을 검사결과 양성으로 판단할 확률은 0.01이다. 통계적으로 1000명 당 5명 꼴로 암에 걸린다고 할 때 검사결과 양성으로 나타난 사람이 실제 암에 걸렸을 확률은?

예제 44

어느 대학교 졸업생 중 남학생의 25%와 여학생의 10%는 은행에 취업했다. 여자 졸업생은 전체 졸업생의 60%이다. 임의로 한 학생을 선택하였더니 은행에 취업한 학생이었다면 그 학생이 여학생일 확률은?

예제 45

어느 도시에서 남성의 4%와 여성의 1%가 1.8m보다 크다. 또 여성은 전체도시 인구의 60%이다. 임의로 한 사람을 선택하였더니 1.8m보다 크다면 그 사람이 여성일 확률은?

제 2 장 확률변수와 확률분포

Def (6): 확률변수

표본공간의 각 원소에 하나의 실수값을 대응시키는 함수를 **확률변수**라 한다.

(1) 이산형 확률변수(discrete random variable)

⇨ X들의 집합이 셀 수 있을 때

(2) 연속형 확률변수(continuous random variable)

⇨ X가 연속적인 구간 내의 값을 취하면

$\Rightarrow$

예제 46

두 개의 주사위를 던져서 나오는 두 눈의 차를 확률변수 X라 할 때 확률변수 X 각각의 확률을 구하여라.

Def (7) : 이산확률변수 X의 확률함수(=확률질량함수)

이산확률변수 X가 취할 수 있는 값 x_1, x_2, $\cdots$, x_n의 각각에 대하여 확률 $P(X=x_1)$, $P(X=x_2)$, $\cdots$, $P(X=x_n)$를 대응시켜주는 관계를 X의 **확률질량함수**라 한다. 그 표시는 $f(x)=P(x_i)$로 한다.

※ 확률질량함수의 성질

① 모든 x에 대하여 $f(x_i) \geq 0$

② $\sum_{i=1}^{\infty} f(x_i)=1$

③ $P(a \leq X \leq b)=\sum_{a \leq x_i \leq b} f(x_i)$

※ 확률분포표

X	x_1	x_2	x_3	$\cdots$	합
$f(x_i)=P(X=x_i)$	$f(x_1)$	$f(x_2)$	$f(x_3)$	$\cdots$	1

예제 47

동전을 3번 던질 때 앞면이 나오는 횟수를 X라 하면 X의 확률분포를 구하여라.

예제 48

어떤 가게에서 판매되는 8대의 전화기 중 3대는 불량품이다. 이 가게에서 임의로 두 대의 전화기를 구입할 때 불량품의 개수의 확률분포를 구하여라.

예제 49

H자동차 대리점에서 판매되는 차의 절반은 듀얼 에어백이 장착되어 있다. 이 대리점에서 판매될 4대의 자동차 가운데 듀얼 에어백이 장착된 차의 수의 확률분포를 구하여라.

예제 50

동전을 3번 던질 때 앞면이 나오는 횟수를 X라 할 때 $Y=X-E(X)$의 확률분포를 구하여라.

Def (8): 이산확률변수 X의 누적분포함수 $F(x)$

확률분포 $f(x)$를 가지는 이산형 확률변수 X의 **누적분포함수**

$$F(x) = P(X \le x) = \sum_{x_i \le x} P(x_i) \text{ 이다. } \quad (\text{단, } -\infty < x < \infty)$$

$\Rightarrow$ 누적분포는 주어진 확률변수가 취할 수 있는 값 뿐만 아니라 모든 실수에서 정의됨에 유의!

① $F(-\infty) = 0$, $F(\infty) = 1$, $0 \le F(x) \le 1$

② $P(a < X \le b) = F(b) - F(a)$

예제 51

두 개의 동전을 던져서 앞면이 나오는 횟수를 X라 할 때 누적분포함수를 구하고 그 그래프를 그려라.

예제 52

A, B, C 세 명의 수험생에게 수험표를 ABC 순서로 나누어 줄 때 자기 수험표를 받게 되는 수험생 수를 X라 할 때 확률변수 X의 누적분포를 구하여라.

예제 53

예제 49에서 확률변수 X의 누적분포를 구하여라. 또, $F(x)$를 이용하여 $f(2)=\frac{3}{8}$ 임을 증명하여라.

예제 54

X의 분포가 다음과 같을 때 누적분포함수 F의 그래프를 그려라.

x_i	-2	3	5
$f(x_i)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

예제 55

X의 누적분포함수 $F(x)=\begin{cases} 0 & , \ (x<0) \\ \frac{1}{2} & , \ (0 \le x < 1) \\ \frac{3}{5} & , \ (1 \le x < 2) \\ \frac{4}{5} & , \ (2 \le x < 3) \\ 1 & , \ (x \ge 3) \end{cases}$ 일 때 다음을 구하여라.

(1) $P(X=0)$

(2) $P(X=1)$

(3) $P(X=2)$

(4) $P(X=3)$

예제 56

어떤 대리점에서 판매되는 TV는 7대 중 2대가 불량품이라고 한다. 이들 TV 중 임의로 3대를 구입할 때 불량 TV가 포함되는 수를 확률변수 X라 할 때

(1) X의 확률분포를 구하여라.

(2) X의 누적분포함수를 구하여라.

(3) 누적분포함수 $F(x)$를 이용하여 다음을 구하여라.

① $P(0 < X \leq 1)$

② $P(0 < X \leq 2)$

③ $P(1 < X \leq 2)$

Def (9): 확률밀도함수(probability density function)

연속확률변수 X에 대하여

① 모든 $x \in R$에 대하여 $f(x) \geq 0$

② $\int_{-\infty}^{\infty} f(x)\,dx = 1$

③ $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$

을 만족하는 $f(x)$를 X의 **확률밀도함수**라 한다.

$\Rightarrow$

예제 57

연속확률변수 X의 확률밀도함수가

$f(x) = \begin{cases} c(1+x)^{-2}, & (x > 0) \\ 0 & , (x \leq 0) \end{cases}$ 일 때 $P(1 \leq X \leq 2)$를 구하여라.

예제 58

연속확률변수 X의 확률밀도함수가

$f(x) = \begin{cases} cx^2, & (0 \leq x \leq 3) \\ 0, & (\text{그 외}) \end{cases}$ 일 때 $P(1 \leq X \leq 2)$를 구하여라.

Def (10)

확률밀도함수가 $f(x)$ 인 연속형 확률변수 X의 **누적분포함수**

$$F(x)=P(X\le x)=\int_{-\infty}^{x} f(t)\,dt,\quad -\infty<x<\infty$$ 이다.

$$\Rightarrow\ P(a\le X\le b)=P(a<X\le b)=P(a\le X<b)=P(a<X<b)$$
$$=\int_{-\infty}^{b} f(t)\,dt-\int_{-\infty}^{a} f(t)\,dt=F(b)-F(a)$$

예제 59

확률밀도함수 $f(x)=\begin{cases} 1\,, & (0\le x\le 1) \\ 0\,, & (x<0 \text{ 또는 } x>1) \end{cases}$ 일 때 누적분포함수 $F(x)$ 를 구하고 $P\left(\frac{1}{4}\le X<\frac{1}{2}\right)$ 를 F를 이용하여 구하여라.

예제 60

확률밀도함수 $f(x)=\begin{cases} \frac{1}{3}x^2, & (-1 < x < 2) \\ 0, & (\text{그 외}) \end{cases}$ 일 때 누적분포함수 F를 구하고 이를 이용하여 $P(0 < X \le 1)$를 구하여라.

예제 61

확률밀도함수 $f(x)=\begin{cases} kx\,, & (0 \le x \le 5) \\ 0\,, & (\text{그 외}) \end{cases}$ 일 때

(1) 누적분포함수 F를 구하라.

(2) F를 이용하여 다음 확률을 구하여라.

$P(1 \le X \le 3)$

$P(2 \le X \le 4)$

$P(X \le 3)$

예제 62

확률밀도함수 $f(x)=\begin{cases} 2e^{-2x}\,, & (x \ge 0) \\ 0\,, & (x < 0) \end{cases}$ 일 때 $P(X > 2)$를 구하여라.

Def (11) : 이산확률변수 X와 Y의 결합확률분포(=결합확률질량함수)

두 개의 이산확률변수 X와 Y가 각각 $x_1,\ x_2,\ \cdots$ 와 $y_1,\ y_2,\ \cdots$ 의 값을 취할 때 $P(X=x,\ Y=y)=f(x,\ y)$를 만족하는 $f(x,\ y)$를 X와 Y의 **결합확률분포** 또는 **결합확률질량함수**라 한다.

※ 이산형 결합확률질량함수의 성질

① $f(x,\ y)\geq 0$ ② $\sum_x \sum_y f(x,\ y)=1$

③ $x,\ y$ 평면 상의 어떤 영역 A에 대하여

$P[(x,\ y)\in A]=\sum\sum_A f(x,\ y)$ 이다.

예제 63

3개의 검은 구슬, 2개의 붉은 구슬, 3개의 흰 구슬이 들어 있는 상자에서 임의로 2개의 구슬을 꺼낼 때 검은 구슬의 개수를 X개, 붉은 구슬의 개수를 Y개라 하면 X와 Y의 결합확률분포를 구하여라. 또, $A=\{(x,\ y)\,|\,x+y\leq 1\}$일 때 $P[(X,\ Y)\in A]$를 구하여라.

예제 64

동전을 2개 던질 때 앞면이 나오는 개수를 확률변수 X라 하고

확률변수 $Y=\begin{cases} 0, & X=0,\ 2 \\ 1, & X=1 \end{cases}$ 로 정의할 때 X와 Y의 결합확률분포를 구하여라.

예제 65

볼펜이 A사의 제품 3개, B사 제품 2개, C사 제품 3개가 들어있는 상자에서 임의로 4개의 볼펜을 선택한다. 이 때 A사 제품 개수를 X, B사 제품 개수를 Y라 할 때 X와 Y의 결합확률분포를 구하여라. 또, A의 영역이 $\{(x,\ y)\,|\, x+y \le 2\}$일 때 $P\{(X,\ Y)\in A\}$를 구하여라.

Def (12) : 연속확률변수 X와 Y의 결합확률밀도함수(or 결합밀도함수)

두 개의 연속확률변수 X와 Y에 대하여 다음을 만족하는 함수 $f(x,\ y)$를 **결합확률밀도함수**라 한다.

① 모든 $(x,\ y)$에 대하여 $f(x,\ y) \geq 0$

② $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,\ y)\,dydx = 1$

③ $P[(x,\ y) \in A] = \iint_A f(x,\ y)\,dydx$

(여기서, A는 xy평면의 임의의 영역)

$\Rightarrow$

ex) $f(x,\ y) = e^{-(x+y)}$ $(x \geq 0,\ y \geq 0)$이고, $Z = X + Y$일 때 $P(Z \leq 1)$를 구하여라.

예제 66

연속확률변수 X와 Y의 결합밀도함수가

$$f(x,\ y)=\begin{cases} \dfrac{2}{5}(2x+3y), & (0 \le x \le 1,\ 0 \le y \le 1) \\ 0 & ,\ (\text{그 외}) \end{cases}$$ 일 때

(1) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,\ y)\,dydx = 1$ 임을 보여라.

(2) $A = \left\{(x,\ y)\ \middle|\ 0 < x < \dfrac{1}{2},\ \dfrac{1}{4} < y < \dfrac{1}{2}\right\}$ 일 때 $P[(X,\ Y) \in A]$ 를 구하여라.

예제 67

$(X,\ Y)$의 결합밀도함수가 $f(x,\ y)=\begin{cases} xye^{-x-y}, & (x>0,\ y>0) \\ 0, & (\text{그 외}) \end{cases}$ 일 때 $P(X \geq 2Y)$를 구하여라.

Def (13) : 주변확률분포(marginal probability distribution)

결합확률분포함수 $f(x,\ y)$가 X 또는 Y만의 분포이면

(1) 이산확률변수인 경우

$$f_X(x) = f_1(x) = g(x) = \sum_y f(x,\ y) \Leftarrow \text{열의 합}$$

$$f_Y(x) = f_2(y) = h(y) = \sum_x f(x,\ y) \Leftarrow \text{행의 합}$$

(2) 연속확률변수인 경우

$$f_1(x) = \int_{-\infty}^{\infty} f(x,\ y)\,dy,\ f_2(y) = \int_{-\infty}^{\infty} f(x,\ y)\,dx$$

예제 68

다음 결합확률분포(예제 63 참고)에서 X, Y의 주변확률분포를 각각 구하여라.

$Y \backslash X$	0	1	2	행의 합
0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
1	$\frac{6}{28}$	$\frac{6}{28}$	0	$\frac{12}{28}$
2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
열의 합	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$	1

예제 69

연속확률변수 X와 Y의 결합밀도함수가

$f(x,\ y)=\begin{cases} \dfrac{2}{5}x+\dfrac{3}{10}y\,, & (0 \le x \le 1,\ 0 \le y \le 2) \\ \quad 0 & ,\ (\text{그 외}) \end{cases}$ 일 때 각각의 주변확률분포를 구하여라.

예제 70

연속확률변수 $X,\ Y$의 결합밀도함수가

$f(x,\ y)=\begin{cases} 2\,, & (0 < x < y < 1) \\ 0\,, & (\text{그 외}) \end{cases}$ 일 때 각각의 주변확률분포를 구하여라.

예제 71

연속확률변수 $X,\ Y$의 결합밀도함수가 $f(x,\ y)=\begin{cases} xye^{-x-y}\,, & (x>0,\ y>0) \\ \quad 0 & ,\ (\text{그 외}) \end{cases}$ 일 때 각각의 주변확률분포를 구하여라.

예제 72

$(X,\ Y)$의 결합밀도함수가 $f(x,\ y)=\begin{cases} x^2+\dfrac{xy}{3}, & (0\le x\le 1,\ 0\le y\le 2) \\ 0, & (\text{그 외}) \end{cases}$ 일 때

(1) $X,\ Y$의 각각의 주변확률분포를 구하여라.

(2) $P\left(X>\dfrac{1}{2}\right),\ P(1\le Y\le 3)$를 각각 구하여라.

(3) $P(X+Y\ge 1)$를 구하여라.

Def (14) : 조건부 분포(conditional distribution)

X와 Y를 이산형 또는 연속형 두 확률변수라 하자.

(1) $X=x$로 주어진 확률변수 Y의 조건부 분포는

$$f(y/x)=\frac{f(x,\ y)}{f_1(x)}$$ (단, $f_1(x)$는 X의 주변밀도함수)

(2) $Y=y$로 주어진 확률변수 X의 조건부 분포는

$$f(x/y)=\frac{f(x,\ y)}{f_2(y)}$$ (단, $f_2(y)$는 Y의 주변밀도함수)

※ 연속형의 경우

(1) $$\int_{-\infty}^{\infty} f(y/x)\,dy=\frac{1}{f_1(x)}\int_{-\infty}^{\infty} f(x,\ y)\,dy=\frac{1}{f_1(x)}f_1(x)=1$$

$$\int_{-\infty}^{\infty} f(x/y)\,dy=\frac{1}{f_2(y)}\int_{-\infty}^{\infty} f(x,\ y)\,dx=\frac{1}{f_2(y)}f_2(y)=1$$

(2) $$P(a<Y<b/X=x)=\int_a^b f(y/x)\,dy=\frac{1}{f_1(x)}\int_a^b f(x,\ y)\,dy$$

$$P(a<X<b/Y=y)=\int_a^b f(x/y)\,dx=\frac{1}{f_2(y)}\int_a^b f(x,\ y)\,dx$$

예제 73

두 확률변수 X, Y의 결합확률밀도함수가

$$f(x,\ y)=\begin{cases} 10xy^2, & (0<x<y<1) \\ 0, & (\text{그 외}) \end{cases}$$ 일 때

(1) 조건부 밀도함수 $f(y/x)$를 구하여라.

(2) $P\left(Y>\frac{1}{2}/X=\frac{1}{4}\right)$를 구하여라.

예제 74

두 확률변수 X, Y의 결합확률밀도함수가

$f(x,\ y)=\begin{cases} x^2 e^{-xy-x}, & (x>0,\ y>0) \\ 0, & (\text{그 외}) \end{cases}$ 일 때 조건부확률밀도함수 $f(y/x)$를 구하여라.

예제 75

이변수 결합확률밀도함수가 $f(x,\ y)=\begin{cases} \dfrac{3}{10}xy-\dfrac{3}{10}x^2, & (0 \le x \le y,\ 1 \le y \le 3) \\ 0, & (\text{그 외}) \end{cases}$ 일 때

조건부확률 $P(X>1/2<Y<3)$를 구하여라.

예제 76

이변수 결합확률밀도함수가 $f(x,\ y)=\begin{cases} \frac{1}{4}x+\frac{3}{4}xy^2, & (0<x<2,\ 0<y<1) \\ 0 & ,\ (\text{그 외}) \end{cases}$ 일 때

$f_1(x),\ f_2(y),\ f(x/y)$ 를 구하고 $P\left(\frac{1}{4}<X<\frac{1}{2}\diagup Y=\frac{1}{3}\right)$ 를 구하여라.

예제 77

이변수 결합확률밀도함수가 $f(x,\ y)=\begin{cases} \frac{3}{4}-\frac{1}{8}x-\frac{1}{8}y\ , & (0 \le x \le 2,\ 2 \le y \le 4) \\ 0 & ,\ (\text{그 외}) \end{cases}$

일 때 $P(2 \le Y \le 3 / X=2)$를 구하여라.

예제 78

아래와 같이 결합확률분포가 주어졌을 때 $Y=1$로 주어진 X의 조건부분포를 구하고 또, $P(X=0/Y=1)$를 구하여라.

Y \ X	0	1	2	$f_2(y)$
0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
1	$\frac{6}{28}$	$\frac{6}{28}$	0	$\frac{12}{28}$
2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
$f_1(x)$	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$	1

Def (15) : 통계적 독립

X와 Y를 결합확률분포 $f(x,\ y)$와 주변확률분포 $f_1(x)$, $f_2(y)$를 가지는 확률변수라 할 때 모든 $(x,\ y)$에 대하여 $f(x,\ y)=f_1(x)f_2(y)$가 성립하면 확률변수 X와 Y는 **통계적으로 독립**이라고 한다.

$\Rightarrow$ 일반적으로 $f(x_1,\ x_2,\ \cdots,\ x_n)$을 확률변수 $X_1,\ X_2,\ \cdots,\ X_n$의 결합확률분포라 할 때 모든 $(x_1,\ x_2,\ \cdots,\ x_n)$에 대하여 $f(x_1,\ x_2,\ \cdots,\ x_n)=f_1(x_1)f_2(x_2)\ \cdots\ f_n(x_n)$이 성립하면 확률변수 $X_1,\ X_2,\ \cdots,\ X_n$는 서로 통계적 독립이다.

예제 79

서로 독립인 세 개의 확률변수 $X_1,\ X_2,\ X_3$의 공통된 확률밀도함수가

$f(x)=3x^2,\ 0<x<1$일 때 $P\left(X_1 \le \frac{1}{2},\ X_2 \le \frac{1}{2},\ X_3 \le \frac{1}{2}\right)$를 구하여라.

예제 80

서로 독립인 세 개의 확률변수 X_1, X_2, X_3의 공통된 확률밀도함수가

$f(x)=\begin{cases} e^{-x}, & (x>0) \\ 0, & (\text{그 외}) \end{cases}$ 일 때 $P(X_1<2,\ 1<X_2<3,\ X_3>2)$를 구하여라.

예제 81

확률변수 X와 Y의 결합확률밀도함수가 $f(x,\ y)=\dfrac{xy^2}{30}$, $x=1,\ 2,\ 3,\ y=1,\ 2$ 일 때 두 변수 X와 Y의 독립여부를 판정하여라.

예제 82

확률변수 X와 Y의 결합확률밀도함수가

$$f(x,\ y)=\begin{cases} 12xy-12xy^2, & (0<x<1,\ 0<y<1) \\ 0 & , (\text{그 외}) \end{cases}$$

일 때 두 변수 X와 Y의 독립여부를 판정하여라.

예제 83

확률변수 X와 Y의 결합확률밀도함수가

$$f(x,\ y)=\begin{cases} x+y, & (0<x<1,\ 0<y<1) \\ 0 & , (\text{그 외}) \end{cases}$$

일 때 두 변수 X와 Y의 독립여부를 판정하여라.

〈심화 특강 ① : 결합누적분포함수〉

Def : 결합누적분포함수

확률변수 X, Y에 대하여

$F(x,\ y)=P(X\le x,\ Y\le y)$를 X와 Y의 **결합누적분포함**수라 한다.

(1) 이산형인 경우

$$F(x,\ y)=P(X\le x,\ Y\le y)=\sum_{t_1\le x}\sum_{t_2\le y}f(t_1,\ t_2)$$

(2) 연속형인 경우

$$F(x,\ y)=P(X\le x,\ Y\le y)=\int_{-\infty}^{x}\int_{-\infty}^{y}f(t_1,\ t_2)\,dt_2\,dt_1$$

여기서, $\dfrac{\partial^2}{\partial x\partial y}F(x,\ y)=f(x,\ y)$이다.

ex 1)

두 변수 X와 Y의 결합누적분포함수가 $F(x,\ y)=xy,\ 0<x<1,\ 0<y<1$일 때 X와 Y의 결합확률밀도함수 $f(x,\ y)$를 구하고 $P(X^2<Y)$를 구하여라.

ex 2

두 변수 X와 Y의 결합누적분포함수를 $F(x,\ y)$라 할 때
$P(a < X \leq b,\ c < Y \leq d) = F(b,\ d) - F(b,\ c) - F(a,\ d) + F(a,\ c)$임을 증명하여라.

ex 3

두 변수 X와 Y의 결합누적분포함수가 $F(x,\ y) = x^2 y,\ 0 < x < 2,\ 0 < y < 3$일 때
$P(X + Y \leq 1)$를 구하여라.

ex 4

결합확률밀도함수 $f(x,\ y)=\begin{cases} 1\ , & (0<x<1,\ 0<y<1) \\ 0\ , & (\text{기타}) \end{cases}$ 일 때 결합누적분포함수 $F(x,\ y)$를 구하여라.

(풀이)

$$F(x,\ y)=P(X\le x,\ \ Y\le y)=\int_0^x \int_0^y 1\, dt_2\, dt_1 = \int_0^x \left[t_2\right]_0^y dt_1$$

$$=\int_0^x y\, dt_1 = y\left[t_1\right]_0^x = xy \quad (0<x<1,\ 0<y<1)$$

ex 5

결합확률밀도함수 $f(x,\ y,\ z)=\begin{cases} e^{-(x+y+z)}\ , & (0<x,\ y,\ z<\infty) \\ 0\ , & (\text{기타}) \end{cases}$ 일 때 결합누적분포함수 $F(x,\ y,\ z)$를 구하여라.

제 3 장　기댓값과 분산

Def (16) : 기댓값(expectation)

(1) 확률변수 X의 확률밀도함수가 $f(x)$일 때 X의 기댓값은

① 이산형 ⇨ $E(X)=\sum_{x} x f(x)$

② 연속형 ⇨ $E(X)=\int_{-\infty}^{\infty} xf(x)\,dx$

(2) 확률변수 X와 Y의 결합확률밀도함수가 $f(x,\ y)$일 때

① 이산형 ⇨ $\mu_x = E(X)=\sum_{x} x f_1(x)$

$\mu_y = E(Y)=\sum_{y} y f_2(y)$

② 연속형⇨ $\mu_x = E(X)=\int_{-\infty}^{\infty} xf_1(x)\,dx$, $\mu_{g(X)} = E(g(X))=\int_{-\infty}^{\infty} g(x)f_1(x)\,dx$

$\mu_y = E(Y)=\int_{-\infty}^{\infty} yf_2(y)\,dy$, $\mu_{g(Y)} = E(g(Y))=\int_{-\infty}^{\infty} g(y)f_2(y)\,dy$

(2)의 증명

예제 84

X의 확률밀도함수가 $f(x)=\dfrac{\binom{4}{x}\binom{3}{3-x}}{\binom{7}{3}}$, $x=0,\ 1,\ 2,\ 3$일 때 기댓값 $E(X)$를 구하여라.

예제 85
확률변수 X의 확률밀도함수가 $f(x)=xe^{-x}$, $x>0$일 때 X의 기댓값을 구하여라.

예제 86
두 확률변수 X와 Y의 결합확률밀도함수가 $f(x, y)=xye^{-x-y}$, $x>0$, $y>0$일 때 $g(X)=X^2+5$의 기댓값을 구하여라.

Def (17): 확률변수 $g(X,\ Y)$의 기댓값

확률변수 X와 Y의 결합확률밀도함수가 $f(x,\ y)$일 때
확률변수 $g(X,\ Y)$의 기댓값은

① 이산형 ⇨ $\mu_{g(X,\ Y)} = E(g(X,\ Y)) = \sum_x \sum_y g(x,\ y) f(x,\ y)$

② 연속형 ⇨ $\mu_{g(X,\ Y)} = E(g(X,\ Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,\ y) f(x,\ y)\, dx dy$

예제 87
확률변수 X와 Y의 결합확률밀도함수 $f(x,\ y)$가 다음과 같을 때
$g(X,\ Y) = XY$의 기댓값을 구하여라.

Y \ X	0	1	2	3
0	$\frac{4}{84}$	$\frac{18}{84}$	$\frac{12}{84}$	$\frac{1}{84}$
1	$\frac{12}{84}$	$\frac{24}{84}$	$\frac{6}{84}$	0
2	$\frac{4}{84}$	$\frac{3}{84}$	0	0

예제 88
X와 Y의 결합확률밀도함수 $f(x,\ y) = e^{-x-y}$, $x > 0$, $y > 0$일 때 $g(X,\ Y) = X + Y$의 기댓값을 구하여라.

예제 89

확률변수 X와 Y의 결합확률밀도함수가 $f(x,\ y)=\dfrac{x(1+3y^2)}{4}$, $0<x<2$, $0<y<1$ 일 때 $g(X,\ Y)=\dfrac{Y}{X}$의 기댓값을 구하여라.

예제 90

두 연속확률변수 X와 Y가 서로 독립이면 $E(XY)=E(X)E(Y)$임을 보이고 $E(g(X)h(Y))=E(g(X))E(h(Y))$임도 보여라.

예제 91

확률변수 X와 Y의 결합확률분포가 다음과 같을 때 $g(X, Y) = XY$의 기댓값을 구하여라.

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

예제 92

결합확률밀도함수가 $f(x, y) = \begin{cases} x+y, & (0<x<1, \ 0<y<1) \\ 0, & (\text{그 외}) \end{cases}$ 일 때

(1) $E(X)$

(2) $E(2X+3Y)$

(3) $E(XY)$

Def (18) : 분산과 표준편차

분산은 분포가 그 평균으로부터 어느 정도 산포되어 있는지의 척도로 $\mathrm{Var}(X) = E\left[(X-\mu)^2\right]$ 으로 정의된다. ($\because$ $\mu = E(X)$)

또 **표준편차**는 $\sigma = \sqrt{\mathrm{Var}(X)}$ 로 정의된다.

① 확률분포 $f(x)$ 에 대해

이산형 ⇨

연속형 ⇨

② 결합확률분포 $f(x,\ y)$ 에 대해

이산형 ⇨

연속형 ⇨

③ $\mathrm{Var}(X) = E\left(X^2\right) - \{E(X)\}^2$

④ $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$

⑤ X와 Y가 독립이면

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

예제 93

X의 확률분포가 다음과 같을 때 Var(X)를 구하여라.

X	0	1	2	3	4
$f(x)$	0.2	0.1	0.3	0.3	0.4

예제 94

확률변수 X의 확률밀도함수가 $f(x)=\frac{4}{3}-\frac{4}{3}x^3,\ 0<x<1$일 때 Var$(X)$를 구하여라.

예제 95

두 확률변수 X와 Y의 결합확률밀도함수가

$$f(x,\ y)=3e^{-3x-y},\ 0<x<\infty,\ 0<y<\infty$$

일 때 Var(X)를 구하고 Var$(3X+5Y)$를 구하여라.

예제 96

다음 결합확률분포에 대하여 다음 질문에 답하여라.

$Y \backslash X$	1	2	3
2	0.24	0.12	0.24
5	0.16	0.08	0.16

(1) X와 Y의 주변확률분포를 구하여라.

(2) $\mathrm{Var}(X)$와 $\mathrm{Var}(Y)$를 구하여라.

(3) $\mathrm{Var}(X+Y)$를 구하여라.

Def (19) : 공분산(Covariance)

두 확률변수 X와 Y가 같이 변하는 정도의 측도로 $(X-\mu_x)(Y-\mu_y)$의 평균으로 정의된다. 즉, $\mathrm{Cov}(X,\ Y)=E[(X-\mu_x)(Y-\mu_y)]$이다.

(1) 이산형 $\Rightarrow$ $\mathrm{Cov}(X,\ Y)=\sum_x\sum_y(x-\mu_x)(y-\mu_y)f(x,\ y)$

연속형 $\Rightarrow$ $\mathrm{Cov}(X,\ Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_x)(y-\mu_y)f(x,\ y)\,dxdy$

(2) $\mathrm{Cov}(X,\ Y)=E(XY)-E(X)E(Y)$, $\mathrm{Cov}(X,\ X)=\mathrm{Var}(X)$

(3) 두 확률변수 X와 Y가 서로 독립이면 $\mathrm{Cov}(X,\ Y)=0$이다.

(4) $\mathrm{Cov}(aX,\ bY)=ab\mathrm{Cov}(X,\ Y)$

$\mathrm{Cov}(X+a,\ Y+b)=\mathrm{Cov}(X,\ Y)$

$\mathrm{Cov}(X,\ aX+b)=a\mathrm{Var}(X)$

예제 97

두 확률변수 X, Y의 결합확률밀도함수가

$f(x,\ y)=\begin{cases} 3, & (0<x<y,\ 0<y<1) \\ 0, & (\text{그 외}) \end{cases}$ 일 때 X와 Y의 공분산을 구하여라.

예제 98

다음 결합확률분포표에서 X와 Y의 공분산을 구하여라.

Y \ X	0	1	2	$f_2(y)$
0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
1	$\frac{6}{28}$	$\frac{6}{28}$	0	$\frac{12}{28}$
2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
$f_1(x)$	$\frac{10}{28}$	$\frac{15}{28}$	$\frac{3}{28}$	1

예제 99

두 확률변수 X와 Y의 결합확률분포가 $f(x,\ y)=\begin{cases} 4xy, & (0 \le y \le x,\ 0 \le x \le 1) \\ 0\ , & (\text{그 외}) \end{cases}$

일 때 $f_1(x)$와 $f_2(y)$를 구하고 Cov$(X,\ Y)$를 구하여라.

Def (20) : 상관계수(Correlation coeffient)

측정단위에 상관없이 공분산과 같은 역할을 하는 측도로서 공분산을 두 확률변수의 표준편차의 곱으로 나눈 값으로 정의된다.

$$\mathrm{Corr}(X,\ Y)=\frac{\mathrm{Cov}(X,\ Y)}{\sigma_x \cdot \sigma_y},\ -1 \le \mathrm{Corr}(X,\ Y) \le 1$$

※ 상관계수의 분석

$\mathrm{Corr}(X,\ Y)=1$ 이면 X와 Y는 완전히 비례관계

$\mathrm{Corr}(X,\ Y)=-1$ 이면 X와 Y는 완전히 반비례관계

$\mathrm{Corr}(X,\ Y)=0$ 이면 X와 Y는 서로 관련이 없다. (선형종속관계 ×)

(주의) $\mathrm{Cov}(X,\ Y)=0$ 이면 $\mathrm{Corr}(X,\ Y)=0$

즉, X와 Y가 독립이면 $\mathrm{Cov}(X,\ Y)=0$ 이고 $\mathrm{Corr}(X,\ Y)=0$ 이다.

but, 역은 꼭 성립하는 것이 아니다.

※ 상관계수의 성질(참고)

① $\mathrm{Corr}(X,\ Y)=\mathrm{Corr}(Y,\ X)$

② $\mathrm{Corr}(X,\ X)=1$, $\mathrm{Corr}(X,\ -X)=-1$

③ $\mathrm{Corr}(aX+b,\ cY+d)=\begin{cases} \mathrm{Corr}(X,\ Y) : ac>0 \text{일 때} \\ -\mathrm{Corr}(X,\ Y) : ac<0 \text{일 때} \end{cases}$

예제 100

두 확률변수 X, Y에서 $\sigma_x=2.28$, $\sigma_y=5.73$, $\mathrm{Cov}(X,\ Y)=9.38$ 일 때 상관계수를 구하여라.

예제 101

다음과 같이 주어진 결합확률분포에서 X와 Y의 공분산과 상관계수를 구하여라.

$Y \backslash X$	0	1	2	3	$f_2(y)$
0	0	0	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{2}{8}$
1	0	$\frac{2}{8}$	$\frac{2}{8}$	0	$\frac{4}{8}$
2	$\frac{1}{8}$	$\frac{1}{8}$	0	0	$\frac{2}{8}$
$f_1(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

예제 102

두 확률변수 X, Y의 결합밀도함수가 $f(x,\ y) = \frac{1}{10}x + \frac{1}{30}y,\ x = 1,\ 2,\ y = 1,\ 2$ 로 주어졌을 때 X와 Y의 공분산과 상관계수를 구하여라.

예제 103

두 확률변수 X와 Y의 결합확률밀도함수가

$f(x,\ y)=\dfrac{3}{2}x^2+\dfrac{3}{2}y^2,\ 0\le x\le 1,\ 0\le y\le 1$일 때

(1) X와 Y는 서로 독립인지 판정하여라.

(2) $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, $\mathrm{Cov}(X,\ Y)$를 구하여라.

(3) $\mathrm{corr}(X,\ Y)$를 구하여라.

Def (21) : 조건부 기대값

(1) 조건부 확률변수 $Y/X=x$ 의 조건부 기댓값은

이산형 ⇨ $E(Y/X=x)=\sum_{y} yf(y/x)$

연속형 ⇨ $E(Y/X=x)=\int_{-\infty}^{\infty} yf(y/x)\,dy$

(2) 조건부 확률변수 $X/Y=y$ 의 조건부 기댓값은

이산형 ⇨ $E(X/Y=y)=\sum_{x} xf(x/y)$

연속형 ⇨ $E(X/Y=y)=\int_{-\infty}^{\infty} xf(x/y)\,dx$

※ X와 Y가 독립인 경우

이산형 ⇒ $E(Y/X=x)=\sum_{y} yf_2(y)=E(Y)$, $E(X/Y=y)=\sum_{x} xf_1(x)=E(X)$

연속형 ⇒ $E(Y/X=x)=\int_{-\infty}^{\infty} yf_2(y)\,dy=E(Y)$, $E(X/Y=y)=\int_{-\infty}^{\infty} xf_1(x)\,dx=E(X)$

예제 104

두 확률변수 X와 Y의 결합확률밀도함수가 $f(x,\ y)=2$, $0<x<y<1$ 일 때 $Y=y$가 주어졌을 때 X의 조건부 기댓값을 구하여라.

예제 105

X와 Y의 결합확률밀도함수 $f(x,\ y)=x^2e^{-xy-x}$, $0<x<\infty$, $0<y<\infty$ 일 때 $E(Y/X=x)$를 구하여라.

Def (22) : 조건부 분산

$X=x$로 주어졌을 때 확률변수 Y의 **조건부 분산**은

$$\mathrm{Var}(Y/x)=E\left(Y^2/x\right)-\{E(Y/x)\}^2$$ 이다.

예제 106

$f(y/x)=xe^{-xy}$, $0<x<\infty$, $0<y<\infty$, $E(Y/x)=\dfrac{1}{x}$ 일 때 $\mathrm{Var}(Y/x)$를 구하여라.

예제 107

두 연속확률변수 X, Y에 대하여 $E[E(Y/X)]=E(Y)$임을 증명하여라.

예제 108

두 연속확률변수 X, Y에 대하여

$\mathrm{Var}(Y)=E[\mathrm{Var}(Y/X)]+\mathrm{Var}[E(Y/X)]$ 임을 증명하여라.

Thm (12): 마코프 부등식(Markov's Inequality)

X가 음이 아닌 값을 취하고 임의의 상수 $a>0$에 대하여

$P(X \ge a) \le \dfrac{E(X)}{a}$가 성립한다.

$\Rightarrow$

※ 일반적으로 실함수 $u(x)>0$일 때 임의의 상수 $a>0$에 대하여

$P(u(x) \ge a) \le \dfrac{E(u(x))}{a}$이 성립한다.

예제 109

K은행은 고객이 도착하여 평균 5분만 기다린다고 홍보하고 있다. 어떤 고객은 1시간만 기다릴 수 있는 여유가 있다. 이 고객이 1시간 이상 기다릴 확률을 구하여라.

예제 110

확률변수 X의 기댓값이 4일 때 확률 $P(X \ge 10)$의 상한을 구하여라.

Thm (13): 체비셰프 부등식(Chebyshev's Inequality)

확률변수 X가 평균(μ)으로부터 표준편차(σ)의 c배 범위내의 값을 취할 확률은 적어도 $1-\dfrac{1}{c^2}$ 이다. 즉, $P(\mu-c\sigma < X < \mu+c\sigma) \geq 1-\dfrac{1}{c^2}$ 이다.

$\Rightarrow$

예제 111

확률변수 X의 평균 $\mu=7$이고 분산 $\sigma^2=9$일 때 다음을 구하여라.

(1) $P(-5 < X < 19)$

(2) $P(|X-7| \geq 6)$

예제 112

H 자동차 회사에서 한 주 동안 생산되는 자동차의 댓수는 평균 500대로 알려져 있다. 한 주의 생산되는 자동차 댓수의 분산이 100일 때 $P(|X-500| \le 100)$을 구하여라. (단, 확률변수 X는 생산 자동차 댓수)

예제 113

확률변수 X의 기댓값이 μ이고 분산이 σ^2일 때 확률변수 X가 $\mu - k\sigma$와 $\mu + k\sigma$사이일 확률이 적어도 0.9일 k값을 체비셰프 부등식으로 구하여라.

예제 114

확률변수 X의 평균이 $\mu = 24$이고 분산 $\sigma^2 = 16$이라 할 때 확률 $P(16 \le X \le 32)$을 구하여라.

제 4 장　　이산형 확률분포

Def (23) : 베르누이 분포

어떤 시행의 결과가 성공 또는 실패 중 하나로 나타나고 성공일 확률을 p라 할 때 그 결과가 성공이면 확률변수 X가 1을 갖고 실패이면 0을 갖는 X를 **베르누이 확률변수**라 하고 그 확률밀도함수는 $f(1)=p$, $f(0)=1-p$로 주어진다.

※ 베르누이 확률변수 X의 기댓값과 분산

$E(X)=p$, $E(X^2)=p$, $\mathrm{Var}(X)=pq$ (단, $p+q=1$)

예제 115

주사위를 한 번 던져 홀수가 나오면 $X=0$, 짝수가 나오면 $X=1$이라 할 때 X의 기댓값과 분산을 구하여라.

Def (24) : 이항분포

성공확률이 p인 베르누이 시행이 n번 반복되었을 때
성공횟수(즉, n개의 베르누이 변수의 합)를 확률변수 X라 할 때 X를
이항확률변수라 한다. 이 때, X의 확률밀도함수

$$f(x)=f(x:n,\ p)=\binom{n}{x}p^x q^{n-x} \quad (x=0,\ 1,\ \cdots,\ n)$$

이며 $X \sim B(n,\ p)$로 표시한다.

※ 이항분포의 기댓값과 분산
$E(X)=np,\ \mathrm{Var}(X)=npq \quad (p+q=1)$

예제 116
특정 질병에 걸린 실험용 쥐가 자연치유될 확률이 0.4라고 한다.
임상실험에서 10마리의 쥐가 그 질병에 걸렸을 때
(1) 3마리의 쥐가 자연치유될 확률을 구하여라.

(2) 적어도 한 마리의 쥐가 자연치유될 확률을 구하여라.

(3) 2마리 이하의 쥐가 자연치유될 확률을 구하여라.

예제 117

어떤 제품을 검사했을 때 양품일 확률이 $\frac{3}{4}$ 이라 한다. 이 제품 4개를 검사했을 때 2개가 양품일 확률은?

예제 118

4지 선다형 문제를 임의로 20문제 풀 때 7개를 맞춘 확률은?

예제 119

$X \sim B(n,\ p) = P(X) = {}_{10}C_x (0.4)^x (0.6)^{10-x}$ 로 주어졌을 때 X의 평균과 분산을 구하여라.

예제 120

$X \sim B(n,\ p) = P(X) = {}_{20}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{20-x}$ 일 때 X의 평균과 분산을 구하여라.

Def (25) : 다항분포

각 시행에서 p_1, p_2, p_3의 확률로 3개의 결과 E_1, E_2, E_3 중 어느 하나가 발생한다면 n번의 독립시행에서 각각 E_1, E_2, E_3의 발생횟수를 나타내는 확률변수를 X_1, X_2, X_3라 할 때 X의 결합확률밀도함수는

$$f(x_1,\ x_2,\ x_3) = \binom{n}{x_1,\ x_2,\ x_3} p_1^{x_1} p_2^{x_2} p_3^{x_3} \ (\text{단},\ x_1+x_2+x_3=n,\ p_1+p_2+p_3=1)$$

예제 121

주사위를 4번 던질 때 1 or 2가 1번, 3 or 4 or 5가 2번, 6의 눈이 1번 나올 확률을 구하여라.

예제 122

A도시 시민은 출근길 교통 수단으로 자가용을 이용할 확률은 $\frac{2}{9}$, 버스는 $\frac{1}{6}$, 지하철은 $\frac{11}{18}$로 알려져 있다. 임의의 6명의 시민이 출근길 교통수단을 이용했을 때 자가용 2명, 버스 1명, 지하철 3명일 확률을 구하여라.

Def (26) : 포아송 분포

확률변수 X가 다음과 같은 확률밀도함수를 가질 때 이를 모수(=단위당 평균발생횟수)가 λ인 **포아송확률변수**라 하고 $X \sim \mathrm{POI}(\lambda)$로 표기한다.

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} = P(x\,;\,\lambda),\quad x = 0,\ 1,\ 2,\ \cdots$$

이고 λ는 평균발생횟수, $e = 2.71828\cdots$ 이다.

포아송 분포 $\Rightarrow$ 주어진 단위시간 또는 단위공간 내에 발생하는 어떤 사건의 횟수를 확률변수로 하는 분포

예제 123

KT 전화국의 어느 전화교환시스템에서는 매분마다 평균 2건의 통화가 이루어진다고 한다. 통화횟수가 포아송분포를 따른다면 3분 동안에 5건 이상의 통화가 이루어질 확률을 구하여라.

예제 124

부산 신항만에는 하루에 평균 10척의 콘테이너선이 도착한다고 알려져 있다. 그런데 이 항만의 시설로는 최대 15척의 콘테이너선만 처리할 수 있다고 한다. 어느날 도착한 콘테이너선이 다른 항만으로 보내질 확률은?

예제 125

어느 도시의 게임 중독자는 평균 1000당 1명이라고 알려져 있다. 7000명을 임의로 조사했을 때 게임중독자가 5명 이하일 확률을 구하여라.

예제 126

어떤 사람은 하루에 핸드폰으로 평균 5번 전화를 받는다고 한다. 어느 특정한 날에 이 사람이 4번 전화를 받을 확률은?

예제 127

어떤 사람은 하루에 핸드폰으로 평균 10회 문자서비스를 받는다고 한다. 어느 특정한 날에 이 사람이 5번의 문자서비스를 받을 확률은?

Thm (14) : 포아송분포의 평균과 분산

$E(X) = \mathrm{Var}(X) = \lambda$ 이다.

$\Rightarrow$ **(증명)**

Thm (15) : 포아송 분포에 의한 이항분포의 근사

확률변수 X가 $B(n,\ p)$분포를 따를 때 p가 상당히 작을 때 n이 한없이 커지고 $np = \lambda$를 만족하면 $x = 0,\ 1,\ 2,\ \cdots$ 에 대하여

$$\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} = \frac{e^{-\lambda}\lambda^x}{x!}$$

이 성립한다.

$\Rightarrow$ **(증명)**

예제 128

어느 휴대폰 회사에서 생산휴대폰의 불량률은 1.5%라 알려져 있다. 임의로 120개의 생산휴대폰을 조사했을 때 불량품이 하나도 없을 확률을 구하여라.

예제 129

어느 자동차 대리점에 1시간 동안 방문하는 고객수는 평균 7인 포아송 분포를 이룬다고 알려져 있다.

(1) 2시간 동안 10명을 초과하는 고객이 방문할 확률을 구하여라.

(2) 2시간 동안 6명의 고객이 방문할 확률을 구하여라.

Def (27) : 초기하분포(hypergeometric distribution)

k개의 성공과 $N-k$개의 실패로 구성된 크기 N인 유한모집단에서 크기 n인 확률표본을 취할 때 성공의 개수를 X라 할 때 이를 **초기하확률변수**라 한다. X의 확률밀도함수

$$f(x)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},\ 0\le x\le k,\ 0\le n-x\le N-k \quad \text{이다.}$$

예제 130

10개 부품이 들어 있는 상자 속에 3개는 불량품이다. 이 상자에서 두 개의 부품을 임의로 추출하여 검사할 때

(1) 2개 모두 불량품일 확률은?

(2) 1개가 불량품일 확률은?

예제 131

100개의 제품 중에 12개가 불량품이다. 100개 중에서 임의로 10개를 뽑았을 때 불량품이 3개일 확률을 구하여라.

예제 132

한 상자 속에 100개의 제품이 들어있는데 여기서 5개를 뽑아 검사하여 불량품이 1개 이하이면 이 상자는 합격이라 판정한다. 만일 이 상자에 실제 불량품이 10개 들어있다고 할 때 합격판정을 받을 확률을 구하여라.

Thm (15) : 초기하분포의 평균과 분산

$E(X) = \dfrac{nk}{N}$, $\mathrm{Var}(X) = \left(\dfrac{N-n}{N-1}\right) n \cdot \dfrac{k}{N}\left(1-\dfrac{k}{N}\right)$ 이다.

$\Rightarrow$

예제 133

120개의 부품이 들어있는 상자 속에 6개는 불량품이다. 이 상자에서 임의로 3개를 추출하였을 때 불량품의 수를 X라 할 때 X의 기대치와 분산을 구하여라.

예제 134

3명의 수학교수와 5명의 생물학교수 중 임의로 5명을 뽑아 위원회를 만들 때 그 위원회에 수학교수의 수를 확률변수 X라 할 때

(1) 확률분포를 구하여라.

(2) 기댓값과 분산을 구하여라.

예제 135

상자 속에 10개의 제품이 들어있는데 이 중에 2개는 불량품이다.

(1) 복원추출로 임의로 2개 꺼낼 때 하나만 불량품일 확률을 구하여라.

(2) 비복원추출로 2개 꺼낼 때 하나만 불량품일 확률을 구하여라.

예제 136

전화기의 5000대 가운데 4000대는 발신자 표시 전화기로 알려져 있다. 임의로 10명에게 전화를 걸었을 때 3명이 발신자 확인이 되지 않는 전화기로 받을 확률은?

예제 137

100개의 제품을 한 묶음으로 하는 묶음 속에 10개의 불량품이 들어있다. 임의로 선택된 5개의 제품 중에 불량품의 수를 확률변수 X라 할 때 X의 기댓값과 분산을 구하여라.

Def (28) : 기하분포(geometric distribution)

성공일 확률이 p인 베르누이 시행을 독립적으로 반복할 때 처음 성공할 때까지의 시행횟수를 X라 할 때 이를 **기하확률변수**라 하고 X의 확률밀도함수 $f(x) = g(x\ ;\ p) = pq^{x-1},\ x = 1,\ 2,\ \cdots,\ p+q = 1$이다.

㉾ 어떤 사람이 앞면이 나오도록 동전을 4번 던질 확률 ⇒

예제 138

100개의 부품마다 평균 1개의 불량품이 들어있다고 알려진 공장에서 제품을 하나씩 검사할 때 불량품이 5번째 검사에서 나올 확률은?

예제 139

매주 발행하는 어떤 복권의 당첨확률은 0.001이라 한다. 어떤 사람이 매주 복권을 하나씩 구입할 때 7주째 당첨할 확률을 구하여라.

예제 140

주사위를 1의 눈이 나올 때까지 던질 때 1의 눈이 3번째에 나올 확률을 구하여라.

Thm (16) : 기하분포의 평균과 분산

$$E(X)=\frac{1}{p}, \ \mathrm{var}(X)=\frac{1-p}{p^2}$$

$\Rightarrow$

예제 141

어떤 사람이 기능사 시험에 합격할 확률이 0.25로 알려져 있다. 이 사람이 합격할 때까지 시험을 치는 횟수를 확률변수 X라 할 때 X의 기댓값과 분산을 구하여라.

Def (29) : 음이항 분포(negative binomial distribution)

성공확률이 p인 베르누이 시행을 독립적으로 반복할 때 k번째 성공이 나올 때까지 시행횟수를 X라 할 때 이를 **음이항 확률변수**라 하고 X의 확률밀도함수는 $f(x\ ;\ k,\ p)=\begin{pmatrix} x-1 \\ k-1 \end{pmatrix} p^k q^{x-k}\ (x=k,\ k+1,\ \cdots\ )$

※ $k=1$일 때 음이항분포=기하분포

※ $X_1,\ X_2,\ \cdots\ ,\ X_k$를 독립기하확률변수라 하면 음이항확률변수 $X=X_1+X_2+\ \cdots\ +X_k$이다.

$\Rightarrow$ x번 시행에서 맨 마지막은 무조건 성공이고 이를 제외한 나머지 $(x-1)$번의 시행에서 $(k-1)$번 성공, $(x-1)-(k-1)=(x-k)$번 실패하는 경우의 수는

$$\underbrace{\overbrace{\bigcirc\bigcirc\ \cdots\ \bigcirc}^{k-1\text{개}}\ \overbrace{\times\times\ \cdots\ \times}^{x-k\text{개}}}_{\text{총}(x-1)\text{개}}\ \overset{\text{고정}}{\bigcirc}$$

를 일렬로 배열하는 방법의 수이므로

$\dfrac{(x-1)!}{(k-1)!(x-k)!}=\begin{pmatrix} x-1 \\ k-1 \end{pmatrix}$이다. 그런데 각 경우의 확률은 항상

$p^{k-1}q^{x-k}p=p^k q^{x-k}$ 이다. 따라서, $f(x\ ;\ k,\ p)=\begin{pmatrix} x-1 \\ k-1 \end{pmatrix} p^k q^{x-k}$ 이다.

예제 142

타율이 3할이 어떤 야구선수가 5번째 타석에 2번째 안타를 칠 확률을 구하여라.

예제 143

실험용 쥐가 특정 전염병에 노출시 그 전염병에 걸릴 확률은 0.4로 알려져 있다. 전염병에 노출된 6번째 실험용 쥐가 3번째로 이 전염병에 걸릴 확률은?

예제 144

갑, 을 두 사람이 7번의 게임을 할 때 먼저 4번을 이기는 팀이 우승한다. 어느 쪽 이든지 먼저 4번을 이기면 더 이상 게임은 계속되지 않고 종료된다. 갑이 매 게임에서 이길 확률은 0.7일 때 경기가 5번 게임으로 종료될 확률을 구하여라.

Thm (17) : 음이항분포의 평균과 분산

$E(X)=\dfrac{k}{p}$, $\mathrm{Var}(X)=\dfrac{k(1-p)}{p^2}$ 이다.

제 5 장　연속형 확률분포

Def (30) : 균일분포(Uniform distribution)

확률변수 X의 확률밀도함수가

$$f(x)=\begin{cases} \dfrac{1}{b-a}, & (a<x<b) \\ 0, & (\text{그 외}) \end{cases}$$ 으로 주어질 때

X는 구간 $(a,\ b)$에서 **균일분포**를 갖는다고 한다.

(1) 누적분포함수 $F(x)=\begin{cases} 0, & (x \le a) \\ \dfrac{x-a}{b-a}, & (a<x<b) \\ 1, & (x \ge b) \end{cases}$ 이다.

(2) $E(X)=\dfrac{a+b}{2}$, $\mathrm{Var}(X)=\dfrac{(b-a)^2}{12}$ 이다.

예제 145

어느 버스 정류장에 버스는 10분 간격으로 도착한다고 한다. 어떤 사람이 임의로 이 정류장에 와서 기다리는 시간은 균일분포를 따른다면

(1) 이 사람이 5분 미만 기다릴 확률을 구하여라.

(2) 이 사람이 7분 이상 기다릴 확률을 구하여라.

예제 146

균일 분포 $f(x)=\begin{cases} \dfrac{1}{10}, & (5<x<15) \\ 0, & (\text{그 외}) \end{cases}$ 일 때

(1) 누적분포함수를 구하여라.

(2) 평균과 분산을 구하여라.

예제 147

어느 회사의 세미나실은 5시간을 초과하여 사용할 수 없다. 세미나 시간 X는 구간 $[0,\ 5]$에서 균일 분포를 이룬다면 어떤 세미나가 4시간 이상 계속될 확률을 구하여라.

예제 148

어떤 회사에서 생산되는 강판은 두께가 200mm에서 250mm 사이의 값을 갖는 균일 분포를 이루고 있다. 만약 그 두께가 220mm 미만이면 불량품으로 처리한다.

(1) 생산된 강판이 불량품일 확률을 구하여라.

(2) 강판 두께 X의 평균과 분산을 구하여라.

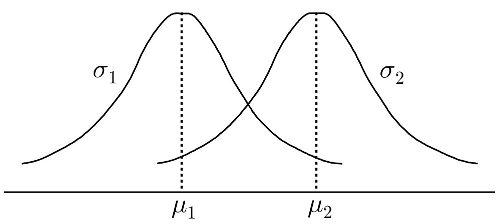
Def (31) : 정규분포(normal distribution)

X의 확률밀도함수가 $f(x\,;\,\mu,\ \sigma)=\dfrac{1}{\sqrt{2\pi}\,\sigma}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $-\infty<x<\infty$

일 때 연속확률변수 X를 평균 μ와 분산 σ^2을 가지는 **정규분포**를 갖는다

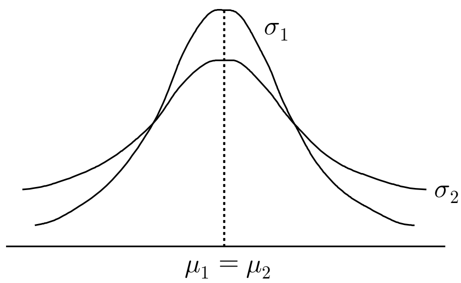
고 한다. $X\sim N(\mu,\ \sigma^2)$로 표기한다. ※ $\displaystyle\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\,\sigma}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx=1$

① $\mu_1\neq\mu_2$, $\sigma_1=\sigma_2$인 정규곡선들



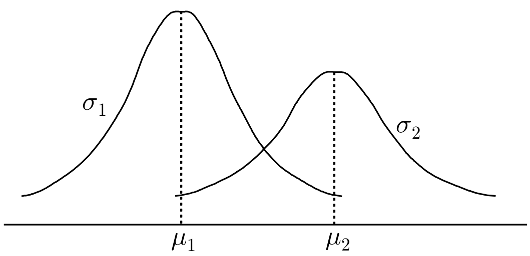
$\Rightarrow$ 평균에 대칭이다.

② $\mu_1=\mu_2$, $\sigma_1<\sigma_2$인 정규곡선들



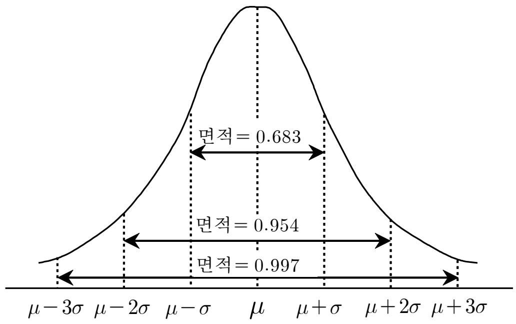
$\Rightarrow$ 편차가 클수록 옆으로 퍼지고 낮아진다.

③ $\mu_1\neq\mu_2$, $\sigma_1<\sigma_2$인 정규곡선들



④ 대표면적

$P(\mu-\sigma<X<\mu+\sigma)=0.683$
$P(\mu-2\sigma<X<\mu+2\sigma)=0.954$
$P(\mu-3\sigma<X<\mu+3\sigma)=0.997$



예제 149

정규확률변수의 기댓값이 μ임을 증명하여라.

예제 150

정규분포의 분산이 σ^2 임을 증명하여라.

Def (32) : 표준정규분포(standard normal distribution)

평균 $\mu=0$ 이고 표준편차 $\sigma=1$ 인 정규분포를 **표준정규분포**라 한다.
표준정규확률변수 Z의 확률밀도함수는

$$f(z\ ;\ 0,\ 1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}},\ -\infty<z<\infty \quad \text{이고}$$

Z의 누적분포함수 $\phi(z)=P(Z\le z)=\int_{-\infty}^{z}f(y\ ;\ 0,\ 1)\,dy$ 이다.

예제 151
다음 확률을 계산하여라.
(1) $P(Z\le 1.5)$

(2) $P(-0.38\le Z\le 1.25)$

Thm (18)-A : 표준화

X의 분포가 $N(\mu,\ \sigma^2)$ 이면 $Z=\dfrac{X-\mu}{\sigma}$ 의 분포는 $N(0,\ 1)$ 이다.

예제 152

확률변수 X가 $N(25,\ 16)$을 따를 때 X가 20과 35사이에 있을 확률 $P(20 \le X \le 35)$를 구하여라.

예제 153

확률변수 X가 $N(25,\ 16)$을 따를 때 $P(|X-25| \le k)=0.95$를 만족하는 상수 k를 구하여라.

예제 154

어느 공장에서 생산되는 부품의 길이는 평균 $370\mathrm{mm}$, 표준편차 $3\mathrm{mm}$의 정규분포를 따를 때 이 제품의 길이가 $376\mathrm{mm}$이상일 확률을 구하여라.

예제 155

어느 회사에서 생산하는 전구는 그 수명이 평균 800시간이고 표준편차가 38.73인 정규분포를 따른다.

(1) 전구의 수명이 778시간과 834시간 사이에 있을 확률을 구하여라.

(2) 수명이 하위 5%에 속하는 제품은 불량품으로 보아 파기한다. 수명이 몇시간 이하면 파기가 되는가?

예제 156

어느 경기에 출전하는 선수의 체중은 평균이 62kg이고 표준편차는 8kg인 정규분포를 따른다. 만일 체중이 50kg에서 74kg까지만 출전이 가능하다면 이 경기에 체중 때문에 출전하지 못할 확률은?

예제 157

어느 대학 통계학 시험 성적이 평균은 74점이고 표준편차 7인 정규분포를 따른다고 한다. 상위 12% 학생에게 A학점을 준다면 A학점을 받기 위한 최소점수는 몇 점인가?

예제 158

400명이 모집정원인 공무원 임용시험에 5000명이 응시하였다. 응시자 전체의 성적분포는 100점 만점에 평균이 55점, 표준편차가 8점인 정규분포를 이루었다. 이 시험에서 모집정원의 120%를 1차 합격자로 선발할 때 1차 합격자의 최저점수를 구하여라. (단, $P(0 \leq Z \leq 1.3) = 0.4040$)

Thm (18)-B : 이항분포의 정규근사

확률변수 X가 $B(n,\ p)$이고 n이 충분히 클 때
$B(n,\ p)$는 $N(np,\ npq)$에 근사한다.
※ 표준화 ⇨ $Z=\dfrac{X-np}{\sqrt{npq}}$ $(\because\ p+q=1)$

※ 일반적으로 $np,\ nq$ 모두 5보다 크면 정규근사가 무난하고 p는 $\dfrac{1}{2}$에 가까운 값일수록 근사가 정확하게 된다.
※ 연속성의 수정을 이용하면 더 정확하게 근사된다.

예제 159
$X\sim B(15,\ 0.4)$일 때 $P(7\le X\le 9)$를 이항분포로 계산한 값과 정규근사해서 구한 값을 비교하여라. 또, 연속성의 수정으로 구한값과도 비교하여라.

Thm (19) : 연속성의 수정을 이용한 이항분포의 정규근사

$X \sim B(n,\ p)$일 때 $\mu = np,\ \sigma^2 = npq$인 정규분포에 근사하면

$P(a \le X \le b) \approx P\left(\dfrac{a-0.5-\mu}{\sigma} \le Z \le \dfrac{b+0.5-\mu}{\sigma}\right)$

$P(X=c) \approx P\left(\dfrac{c-0.5-\mu}{\sigma} < Z < \dfrac{c+0.5-\mu}{\sigma}\right)$이다.

여기서 Z는 $N(1,\ 0)$를 따른다.

예제 160

4지선다형 80문제를 임의로 답을 골랐을 때 정답이 25개에서 30개일 확률을 구하여라.

예제 161

$X \sim B(15,\ 0.4)$일 때 $P(X=4)$를 구하여라.

예제 162

하나의 주사위를 450번 던질 때 3 or 4의 눈이 160번 이상 180번 이하로 나올 확률을 구하여라.

예제 163

K 제약회사의 백신효능은 $\frac{9}{10}$라고 한다. 이 백신을 접종한 사람 중 임의로 100명을 뽑아 조사하였을 때 면역성을 가진 사람이 84명에서 95명 사이일 확률을 구하여라.

Thm (20) : 정규분포를 이루는 확률변수의 합과 차의 분포

서로 독립인 연속 확률변수 X_1, X_2가 각각 $N(\mu_1,\ \sigma_1^2)$, $N(\mu_2,\ \sigma_2^2)$을 따를 때 $X_1 \pm X_2$는 $N(\mu_1 \pm \mu_2,\ \sigma_1^2+\sigma_2^2)$를 따른다.

$Z=\dfrac{(X_1 \pm X_2)-(\mu_1 \pm \mu_2)}{\sqrt{\sigma_1^2+\sigma_2^2}}$ 는 $N(0,\ 1^2)$을 따른다.

$\Rightarrow$

예제 164

인구가 10만인 어떤 도시에서 시정에 대한 여론을 조사하였더니 성인남자 80%와 성인여자 90%가 시정을 지지하였다. 이 도시에서 성인남자 400명과 성인여자 400명을 임의로 뽑았을 때 다음 확률을 구하여라.

(1) 적어도 700명이 시정에 대하여 지지할 확률

(2) 시정에 대한 지지자 중 여자가 남자보다 25명 이상 더 많을 확률

예제 165

어느 학생의 수학 성적은 $N(80,\ 3^2)$을 따르고 물리학 성적은 $N(85,\ 4^2)$을 따를 때 두 과목의 성적의 합이 175점 이하일 확률을 구하여라.

예제 166

빵 한 개의 열량은 평균이 200칼로리, 표준편차는 16칼로리, 우유 1장의 열량은 평균이 80칼로리, 표준편차는 12칼로리의 정규분포를 따른다고 한다.

아침식사에서 300칼로리를 섭취할 수 있는 날은 1년 365일 중 몇 일이 되는지 구하여라.

Def (33) : 지수분포(exponential distribution)

연속형 확률변수 X의 확률밀도함수 $f(x)$가

$$f(x)=f(x\ ;\ \lambda)=\begin{cases} \lambda e^{-\lambda x}, & (x>0,\ \lambda>0) \\ 0\ , & (\text{그 외}) \end{cases}$$ 일 때

X는 **지수분포**를 따른다고 한다.

※ 지수분포는 한 사건에서 다음 사건이 발생할 때까지 걸리는 시간을 확률변수 X라 둔 것으로도 사용된다. 또, 지수분포는 감마분포의 특수한 경우이다.

예제 167

지수분포의 기댓값과 분산을 구하여라.

예제 168

어느 회사의 드라이기는 평균수명이 3년이고 품질보증기간은 1년이라고 한다. 보증기간 내 고장이 나서 새 제품으로 교체해 줄 확률을 구하여라. (단, 제품의 수명시간은 지수분포를 따른다.)

예제 169

어느 회사의 냉장고는 평균 10년 동안 고장이 없다고 한다. 이 냉장고가 20년 동안 고장이 없을 확률을 구하고, 또 5년 이내에 고장날 확률을 구하여라. (단, 고장날 때까지의 시간은 지수분포를 따른다.)

예제 170

어느 은행에 고객이 도착하여 은행업무를 볼 때까지 기다리는 시간은 평균 30분인 지수분포를 따른다고 한다.

(1) 이 고객이 은행업무를 받기까지 40분 이상 기다릴 확률을 구하여라.

(2) 기다리는 시간의 분산을 구하여라.

Thm (21) : 지수분포와 포아송 과정의 관계

모수(=단위당 평균발생횟수)가 λt 인 포아송 분포에서 연속적으로 발생하는 두 사건 사이의 경과시간을 확률변수 X라 하면 X는 $f(x)=\lambda e^{-\lambda x}$ $(x>0)$ 인 지수분포를 따른다.

$\Rightarrow$

예제 171

고장횟수가 포아송 분포를 따르는 어떤 기계는 1개월에 평균 3번의 고장을 일으킨다고 한다.

(1) 이 기계가 고장이 나서 고친 후 2개월 내에는 다시 고장나지 않을 확률을 구하여라.

(2) 기계가 한 번 고장나서 고친 후 다시 고장날 때까지 걸린 시간의 평균과 분산을 구하여라.

Def (34) : 무기억성(memoryless property)

만일 모든 s, $t \geq 0$에 대하여 $P(X > s+t \mid X > t) = P(X > s)$ 이면
음이 아닌 확률변수 X는 **무기억성**을 만족한다고 한다.

※ 지수분포는 무기억성을 가진다.

※ Def (34)의 통계적 의미(지수분포를 전제)

만일 X가 어떤 기계의 수명이라면 주어진 식의 좌변은 기계가 이미 t시간 동안 사용된 상태에 적어도 $s+t$ 시간동안 작동할 확률을 의미하고 우변은 기계가 처음부터 기계가 s시간 동안 작동할 확률을 의미한다.

여기서 "좌변=우변" 이다라는 것은 이미 기계가 t시간 사용했지만 작동될 나머지 시간의 분포는 원래의 수명분포와 같다는 것이다. 즉, 마치 t시간 동안 이미 작동했다는 사실을 기억하지 못하는 것과 같다.

예제 172

어느 자동차 배터리는 수명을 다하기 전에 주행할 수 있는 거리가 평균 5000㎞인 지수분포를 따른다고 한다. 이 자동차로 2000㎞거리 이상을 배터리 교환 없이 갈 수 있는 확률을 구하여라.

Def (35) : 감마함수

$x>0$ 일 때 감마함수 $\Gamma(\alpha)=\int_0^\infty x^{\alpha-1}e^{-x}\,dx$ 로 정의한다.

① $\Gamma(1)=1$

② $\Gamma\left(\frac{1}{2}\right)=\sqrt{\pi}$ $\left(단,\ \int_0^\infty e^{-x^2}dx=\frac{\sqrt{\pi}}{2}\text{ 임을 이용하여라.}\right)$

③ $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$ (단, $\alpha>0$)

④ $\Gamma(n+1)=n!$ (단, n는 자연수)

Def (36) : 감마분포(gamma distribution)

감마분포는 α회의 포아송사건이 발생할 때까지 소요되는 시간의 분포로 X의 확률밀도함수는

$$f(x)=f(x,\ \alpha,\ \beta)=\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\beta}} \qquad \left(\because\ \beta=\frac{1}{\lambda}\right)$$

$$=\frac{1}{(\alpha-1)!\,\beta^{\alpha}}x^{\alpha-1}e^{-\frac{x}{\beta}} \qquad (x>0,\ \alpha>0,\ \beta>0)$$

※ 감마분포와 지수분포의 관계

서로 독립적인 α개의 확률변수 $X_1,\ X_2,\ \cdots,\ X_{\alpha}$ 가 모수 λ인 지수분포를 따를 때 이들의 합인 $X=X_1+X_2+\ \cdots\ +X_{\alpha}$는 감마분포를 따른다.

Thm (22) : 감마분포의 평균과 분산

$E(X)=\dfrac{\alpha}{\lambda}=\alpha\beta,\ \mathrm{Var}(X)=\dfrac{\alpha}{\lambda^2}=\alpha\beta^2$

$\Rightarrow$

예제 173

전화교환기에 도착하는 호출 신호는 분당 평균이 5회인 포아송 과정을 따른다고 한다. 1분 내에 2번의 호출신호가 도착할 확률은?

※ 불완전감마함수표

$F(x\ ;\ \alpha)=\int_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)}$ 일 때

$F(6\ ;\ 5)=0.715$, $F(5\ ;\ 2)=0.96$, $F(3\ ;\ 5)=0.185$

예제 174

어느 제품에 대한 고객의 불만제기 사이의 시간 간격은 (단위:월) $\alpha=2$, $\beta=4$인 감마분포를 따른다고 한다. 최근 품질관리를 철저히 실시하고 난 후 첫 번째 불만이 발생하는데 20개월이 걸렸다. 이것으로부터 품질관리를 철저히 시행한 것이 효과적이었다고 볼 수 있는가?

예제 175

쥐를 이용한 생의학 실험에서 일정량의 독극물에 대한 생존시간(단위:일)은 $\alpha=5$, $\beta=10$인 감마분포를 이룬다고 한다. 어떤 쥐의 생존 시간이 60일 이하일 확률을 구하여라.

예제 176

생의학 실험에서 γ선에 노출된 실험용 쥐의 생존기간 (단위:주)은 $\alpha=5$이고 $\beta=10$인 감마분포를 따른다면

(1) 임의로 선택된 쥐의 평균생존기간과 분산을 구하여라.

(2) 30주 이상 생존할 확률을 구하여라.

제 6 장 확률변수들의 함수

Thm (23) : 이산확률변수변환①

이산확률변수 X의 확률변수가 $f(x)$이고 X와 Y사이에는 $Y=u(X)$라는 일대일 대응관계가 성립해서 $y=u(x)$를 x에 관하여 푼 유일한 $x=\omega(y)$ 일 때 Y의 확률분포 $g(y)=f[\omega(y)]$ 이다.

$\Rightarrow$

예제 177

X의 확률분포 $f(x)=\dfrac{4}{5}\left(\dfrac{1}{5}\right)^{x-1}$, $x=1,\ 2,\ \cdots$ 인 기하분포일 때 $Y=X^2$의 확률분포를 구하여라.

예제 178

X의 확률변수 $f(x)=\begin{cases} \dbinom{3}{x}\left(\dfrac{2}{5}\right)^x\left(\dfrac{3}{5}\right)^{3-x}, & (x=0,\ 1,\ 2,\ 3) \\ \qquad 0 & ,\ (\text{그 외}) \end{cases}$ 일 때 $Y=X^2$의 확률분포를 구하여라.

Thm (24) : 이산확률변수변환②

두 이산확률변수 X_1과 X_2가 결합확률분포가 $f(x_1,\ x_2)$이고 $(x_1,\ x_2)$와 $(y_1,\ y_2)$는 서로 일대일로 대응하여 식 $y_1 = u_1(x_1,\ x_2)$와 $y_2 = u_2(x_1,\ x_2)$를 풀면 유일하게 $x_1 = \omega_1(y_1,\ y_2)$, $x_2 = \omega_2(y_1,\ y_2)$로 될 때 $Y_1 = u_1(X_1,\ X_2)$와 $Y_2 = u_2(X_1,\ X_2)$로 정의되는 새로운 확률변수 Y_1과 Y_2의 결합확률분포는 $g(y_1,\ y_2) = f[\omega_1(y_1,\ y_2),\ \omega_2(y_1,\ y_2)]$가 된다.

예제 179

서로 독립인 확률변수 X_1과 X_2는 각각 5, 6의 모수를 가지는 포아송 분포를 따를 때 새로운 확률변수 $Y_1 = X_1 + X_2$의 확률분포가 $\displaystyle\sum_{y_2=0}^{y_1} \frac{e^{-11}5^{y_1-y_2}6^{y_2}}{(y_1-y_2)!\,y_2!}$ 임을 증명하여라.

예제 180

X_1과 X_2는 다음과 같은 다항분포를 가질 때

$$f(x_1,\ x_2) = \binom{2}{x_1,\ x_2,\ 2-x_1-x_2}\left(\frac{1}{4}\right)^{x_1}\left(\frac{1}{3}\right)^{x_2}\left(\frac{5}{12}\right)^{2-x_1-x_2}$$

$$(x_1 = 0,\ 1,\ 2,\ x_2 = 0,\ 1,\ 2,\ x_1 + x_2 \le 2)$$

$Y_1 = X_1 + X_2$와 $Y_2 = X_1 - X_2$의 결합확률분포를 구하여라.

예제 181

X_1과 X_2의 결합확률분포가 다음과 같을 때 변환된 $Y_1 = X_1 + X_2$와 $Y_2 = X_1 X_2$의 결합확률분포를 구하여라.

(x_1, x_2)	$(0, 0)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
$f(x_1, x_2)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

예제 182

X_1, X_2, X_3의 결합확률분포가 다음과 같을 때

(x_1, x_2, x_3)	$(0, 0, 0)$	$(0, 0, 1)$	$(0, 1, 1)$	$(1, 0, 1)$	$(1, 1, 0)$	$(1, 1, 1)$
$f(x_1, x_2, x_3)$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

변환된 확률변수 $Y_1 = X_1 + X_2 + X_3$와 $Y_2 = |X_3 - X_2|$의 결합확률분포를 구하여라.

Thm (25) : 연속확률변수의 변수변환①

연속확률변수 X의 확률분포가 $f(x)$이고, X와 Y사이에는 일대일 대응 관계가 성립해서 $y=u(x)$를 풀면 유일하게 $x=\omega(y)$로 될 때 Y의 확률분포는 $g(y)=f[\omega(y)]\,|J|$이다.
여기서, $J=\omega'(y)$이며 이를 야코비안이라고 한다.

$\Rightarrow$ ① $y=u(x)$가 증가함수일 때 $(J>0)$

② $y=u(x)$가 감소함수일 때 $(J<0)$

예제 183

연속확률변수 X의 확률분포가 $f(x)=\begin{cases} 2x\,, & (0<x<1) \\ 0\,, & (\text{그 외}) \end{cases}$ 일 때 확률변수 $Y=8X^3$의 확률분포를 구하여라.

예제 184

연속확률변수 X의 확률변수가 $f(x)=\begin{cases} \dfrac{1}{12}x\,, & (1<x<5) \\ 0\,, & (\text{그 외}) \end{cases}$ 일 때 확률변수 $Y=2X+1$의 확률분포를 구하여라.

예제 185

확률변수 X의 확률밀도함수 $f(x)=\begin{cases} \dfrac{32}{(x+4)^3}\,, & (x>0) \\ 0\,, & (\text{그 외}) \end{cases}$ 일 때

(1) $Y=X+4$의 확률밀도함수를 구하여라.

(2) $P(Y>8)$를 구하여라.

Thm (26) : 연속확률변수의 변수변환②

연속확률변수 X_1, X_2의 결합확률변수가 $f(x_1,\ x_2)$이고 $(x_1,\ x_2)$와 $(y_1,\ y_2)$는 서로 일대일로 대응하여 $y_1 = u_1(x_1,\ x_2)$, $y_2 = u_2(x_1,\ x_2)$를 풀면 유용하게 $x_1 = \omega_1(y_1,\ y_2)$와 $x_2 = \omega_2(y_1,\ y_2)$로 될 때 Y_1과 Y_2의 결합확률분포는 $g(y_1,\ y_2) = f[\omega_1(y_1,\ y_2),\ \omega_2(y_1,\ y_2)] \cdot |J|$

여기서, $J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix}$ 이다.

예제 186

연속확률변수 X_1과 X_2는 결합확률분포 $f(x_1,\ x_2) = \begin{cases} 4x_1x_2, & (0 < x_1 < 1,\ 0 < x_2 < 1) \\ 0\ , & (\text{그 외}) \end{cases}$

일 때 $Y_1 = X_1^2$과 $Y_2 = X_1X_2$의 결합확률분포를 구하여라.

예제 187

확률변수 X_1, X_2가 서로 독립이고 각각 평균 $\lambda=1$인 지수분포를 따른다면

(1) $Y_1 = X_1$과 $Y_2 = X_1 + X_2$의 결합확률분포를 구하여라.

(2) (1)에서 주변밀도함수 $f_1(y_1)$과 $f_2(y_2)$를 구하여라.

(3) $Y_1 = X_1 - X_2$, $Y_2 = X_1 + X_2$의 결합확률분포를 구하여라.

예제 188

서로 독립인 확률변수 X_1, X_2가 표준정규분포를 따른다고 한다. 두 확률변수 $Y_1 = X_1 + X_2$와 $Y_2 = \dfrac{X_1}{X_2}$의 결합확률밀도함수를 구하여라.

예제 189

두 확률변수 X, Y의 결합확률밀도함수가 $f(x,\ y) = 2(x+y)$, $0 \le x \le y \le 1$일 때 $Z = X + Y$의 확률밀도함수를 구하여라.

※ $Y = u(X)$ 의 변수변환시 X와 Y가 일대일이 아닌 경우

예제 190

연속확률변수 X의 확률분포 $f(x)$가 $-1 < x < 2$에서 양의 값을 갖고 그 외에서는 0을 갖는다고 한다. $Y = X^2$의 확률분포를 구하여라.

예제 191

Z가 표준정규분포 $N(0,\ 1^2)$을 따를 때 $Y=Z^2$의 분포를 구하여라.

※ 누적분포함수를 이용한 변수변환

예제 192

연속확률변수 X의 확률분포가 $f(x)$일 때 $Y=\dfrac{3}{5}X+27$의 확률분포를 구하여라.

예제 193

연속확률변수 X의 확률변수 $f(x)=\begin{cases} 1\,, & (0 \le x \le 1) \\ 0\,, & (\text{그 외}) \end{cases}$ 일 때 $Y=-2\log X$의 확률분포를 구하여라.

Def (37): 확률변수 X의 원점에 대한 r차 적률(moment)

확률변수 X^r의 기댓값 $E(X^r)$를 확률변수 X의 **원점에 대한 r차 적률**이라 고 하고 이를 $\mu_r{}' = E(X^r)$로 표기한다.

$$\Rightarrow \mu_r{}' = E(X^r) = \begin{cases} \sum_x x^r f(x), & \text{이산형} \\ \int_{-\infty}^{\infty} x^r f(x)\,dx, & \text{연속형} \end{cases}$$

⇒ 확률변수의 적률은 분포의 특징을 설명해 주는 중요한 역할을 한다.
평균, 분산, 왜도, 첨도는 모두 적률의 함수이다.

⇒ $\mu_r = E\left[(X-\mu)^r\right]$는 평균 μ에 대한 r차 중심적률이라 정의한다.

Def (38): 적률생성함수 $M_X(t)$

확률변수 X의 적률생성함수 $M_X(t) = E(e^{tX})$로 정의한다.

$$\Rightarrow M_X(t) = E(e^{tX}) = \begin{cases} \sum_x e^{tx} f(x), & \text{이산형} \\ \int_{-\infty}^{\infty} e^{tx} f(x)\,dx, & \text{연속형} \end{cases}$$

Thm (27): 적률생성함수로 r차 적률 구하기

$$\left.\frac{d^r}{dt^r} M_X(t)\right|_{t=0} = M_X^{(r)}(0) = E(X^r) = \mu_r{}'$$

⇒

예제 194

확률변수 X가 $B(n,\ p)$를 따를 때 X의 적률생성함수를 구하고 이를 이용하여 $\mu = np,\ \sigma^2 = npq$임을 증명하여라.

예제 195

확률변수 X가 모수가 λ인 포아송 분포를 따를 때 X를 적률생성함수를 구하고 이를 이용하여 $\mu = \lambda,\ \sigma^2 = \lambda$임을 증명하여라.

예제 196

확률변수 X가 표준정규분포 $N(0,\ 1^2)$를 따를 때, X의 적률생성함수를 구하여라.

예제 197

확률변수 X가 정규분포 $N(\mu,\ \sigma^2)$를 따를 때 X의 적률생성함수는

$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$ 임을 증명하여라.

Thm (28)

확률변수 X와 Y가 같은 적률생성함수를 가지면 즉, 모든 t에 대해 $M_X(t)=M_Y(t)$이면 두 확률변수는 같은 확률분포를 가진다.

$\Rightarrow$ 확률변수의 분포형태가 적률생성함수에 의하여 유일하게 결정됨을 보여준다. (증명은 학부 통계학 수준을 넘으므로 생략함)

예제 198

(1) $M_{X+a}(t)=e^{at}M_X(t)$ 임을 증명하여라.

(2) $M_{aX}(t)=M_X(at)$ 임을 증명하여라.

예제 199

X가 표준정규분포를 따를 때 $Y=aX+b$의 적률생성함수를 구하여라.

(참고)

$Y=aX+b$일 때 $M_Y(t)=M_{aX+b}(t)=e^{bt}M_X(at)$ 이다.

Thm (29)

서로 독립인 확률변수 X_1, X_2, $\cdots$, X_n의 적률생성함수가 각각 $M_{X_1}(t)$, $M_{X_2}(t)$, $\cdots$, $M_{X_n}(t)$일 때 $Y = X_1 + X_2 + \cdots + X_n$의 적률생성함수 $M_Y(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t)$ 이다.

$\Rightarrow$

예제 200

서로 독립인 두 확률변수 X_1, X_2가 각각 모수가 λ_1, λ_2인 포아송 분포를 따를 때 $Y = X_1 + X_2$의 적률생성함수를 구하여라.

예제 201

X_1, X_2, $\cdots$, X_n이 서로 독립인 베르누이 확률변수일 때 $Y = X_1 + X_2 + \ \cdots \ + X_n$의 적률생성함수를 구하고 Y는 이항분포 $B(n,\ p)$를 따름을 보여라. (단, $P(X_i = 1) = p$, $P(X_i = 0) = q$, $p + q = 1$이다.)

예제 202

서로 독립인 확률변수 X_1, X_2, $\cdots$, X_n이 각각 모수가 λ_i인 포아송 분포를 따른다면 $Y = X_1 + X_2 + \ \cdots \ + X_n$는 모수가 $\sum_{i=1}^{n} \lambda_i$인 포아송 분포를 따름을 보여라.

예제 203

X_1, X_2는 서로 독립이면서 각각 정규분포 $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$을 따를 때 $Y = X_1 + X_2$는 정규분포 $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$를 따름을 보여라.

예제 204

서로 독립인 X_1, X_2, $\cdots$, X_n이 각각 확률밀도함수 $f(x_i) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$, $x > 0$을 가지는 지수분포를 따를 때 $Y = X_1 + X_2 + \cdots + X_n$은 감마분포를 따름을 보여라.

여기서, $\beta = \frac{1}{\lambda}$이고, $M_{X_i}(t) = \frac{1}{1-\beta t}$이고 Z가 감마분포를 따르면 $M_Z(t) = \left(\frac{1}{1-\beta t}\right)^n$이다.

Def (39) : 카이제곱분포

연속확률변수 X의 확률분포가 $f(x\,;v)=\dfrac{1}{2^{\frac{v}{2}}\Gamma\left(\frac{v}{2}\right)}x^{\frac{v}{2}-1}\cdot e^{-\frac{x}{2}}$, $x>0$일 때

X는 자유도 v인 **카이제곱분포**를 따른다고 한다.

$\Rightarrow$

예제 205

연속확률변수 X가 $N(\mu,\ \sigma^2)$을 따를 때 $Y=\dfrac{(X-\mu)^2}{\sigma^2}$은 자유도 1인 카이제곱분포를 따름을 증명하여라.

예제 206

확률변수 X가 가유도 v인 카이제곱분포를 따를 때 X의 적률생성함수는 $M_X(t)=(1-2t)^{-\frac{v}{2}}$ 임을 보여라.

예제 207

자유도 v인 카이제곱분포의 평균과 분산은 각각 v와 $2v$임을 증명하여라.

Thm (30)

확률변수 X_1, X_2, $\cdots$, X_n 이 서로 독립이고 각각 자유도가 v_1, v_2, $\cdots$, v_n 인 카이제곱분포를 따른다면 $Y = X_1 + X_2 + \cdots + X_n$ 은 자유도가 $v = v_1 + v_2 + \cdots + v_n$ 인 카이제곱분포를 따른다.

$\Rightarrow$

Thm (30-1)

확률변수 X_1, X_2, $\cdots$, X_n 이 서로 독립이면서 $N(\mu,\ \sigma^2)$ 을 따른다면 $Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2$ 은 자유도 $v = n$ 인 **카이제곱분포**를 따른다.

$\Rightarrow$

제 7 장 표본 분포

Def (40): 표본평균과 표본분산

$X_1,\ X_2,\ \cdots,\ X_n$ 을 크기 n 인 확률표본이라 할 때

(1) 표본평균 $\overline{X}=\dfrac{1}{n}(X_1+X_2+\ \cdots\ +X_n)=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_i$

(2) 표본분산 $s^2=\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$

예제 208

어느 회사의 탄소음료캔 속에 들어있는 당분의 비율(%)을 알아보기 위해 7개의 캔을 조사한 결과 18, 21, 17, 16, 9, 27, 18을 얻었다. 이 확률표본의 표본평균과 분산을 구하여라.

Thm (31)

s^2 이 크기 n 인 확률표본의 분산이라면

$s^2=\dfrac{1}{n(n-1)}\left[n\displaystyle\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right]$ 으로 쓸 수 있다.

$\Rightarrow$

예제 209

임의로 선정된 6명의 학생들의 수학시험에서 틀린 문제의 개수는 3, 4, 5, 6, 6, 7이었다. 표본분산을 구하여라.

Thm (32) : 표본평균 $\overline{X}$의 평균과 분산

모평균 μ, 모분산 σ^2인 무한모집단 또는 복원추출인 유한모집단으로부터 임의추출한 크기 n인 표본평균 $\overline{X}$의 평균과 분산은

$$E(\overline{X}) = \mu, \quad \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$$ 이다.

$\Rightarrow$

예제 210

$E(s^2) = \sigma^2$ 임을 증명하여라.

Thm (33) : 정규모집단으로부터의 표본평균

정규모집단 $N(\mu,\ \sigma^2)$ 으로부터 크기 n인 표본 $(X_1,\ X_2,\ \cdots\ ,\ X_n)$의

평균 $\overline{X}$의 분포는 $N\left(\mu,\ \dfrac{\sigma^2}{n}\right)$를 따르고

$Z=\dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ 는 표준정규분포 $N(0,\ 1^2)$ 을 따른다.

$\Rightarrow$

예제 211

정규모집단 $N(50,\ 2^2)$ 에서 크기 $n=16$ 인 표본의 평균을 $\overline{X}$라 할 때 $P(\overline{X}\geq 51)$를 구하여라.

예제 212

정규모집단 $N(75,\ 10^2)$ 으로부터 크기 $n=25$ 인 표본의 평균을 $\overline{X}$라 할 때 $P(71<\overline{X}<79)$를 구하여라.

예제 213

어느 자동차 밧데리 회사에서 생산되는 밧데리의 평균수명은 800일이고 표준편차가 40일인 정규분포를 따른다고 한다. 크기 $n=16$으로 랜덤추출한 표본의 평균이 775일 미만일 확률을 구하여라.

예제 214

K 초등학교 2학년의 학생 키는 평균 130cm이고 표준편차는 10cm이라고 한다. 이 학생들을 25명을 한 반으로 할 때 한 반의 키의 평균이 135cm이상일 확률을 구하여라.

예제 215

항공기 부품을 생산하는 회사에서 특정 부품의 직경의 평균은 5mm, 표준편차는 0.1mm이라고 한다. 임의로 추출한 100개의 부품의 평균직경이 5.027이상일 확률을 구하여라.

Thm (34) : 정규확률변수들의 합의 분포

서로 독립인 확률변수 X_1, X_2, $\cdots$, X_n 이 각각 평균이 μ_1, μ_2, $\cdots$, μ_n 이고 분산이 σ_1^2, σ_2^2, $\cdots$, σ_n^2 인 정규분포를 따를 때 그들의 합 $\sum_{i=1}^{n} X_i$ 는 정규분포 $N\left(\sum_{i=1}^{n}\mu_i,\ \sum_{i=1}^{n}\sigma_i^2\right)$ 을 따른다.

$\Rightarrow$

Thm (35) : 두 정규 모집단으로부터의 표본평균 차의 분포

두 정규 모집단이 서로 독립이고 각각의 평균이 μ_1, μ_2 이고 분산이 σ_1^2, σ_2^2 일 때 각 정규 모집단으로부터 임의로 추출된 크기 n_1, n_2 인 두 표본 평균의 차이 $\overline{X_1} - \overline{X_2}$ 의 분포는 정규분포 $N\left(\mu_1 - \mu_2,\ \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$ 을 따른다.

$\Rightarrow$

※ 정규모집단이 아닌 경우라도 어느 정도 근사적으로 $\overline{X_1} - \overline{X_2}$ 는 정규분포 $N\left(\mu_1 - \mu_2,\ \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$ 을 따른다.

예제 216

어떤 부품 한 개를 만드는 데 걸리는 시간 X는 $N(6,\ 2^2)$을 따른다면 부품 10개를 만드는 데 걸리는 시간이 70이상일 확률을 구하여라.

예제 217

두 정규분포 $N(50,\ 3^2)$, $N(40,\ 2^2)$으로부터 각각 크기 $n_1 = 5$, $n_2 = 4$인 표본을 랜덤추출하였을 때 표본평균의 차 $\overline{X_1} - \overline{X_2}$가 8.2보다 작을 확률을 구하여라.

예제 218

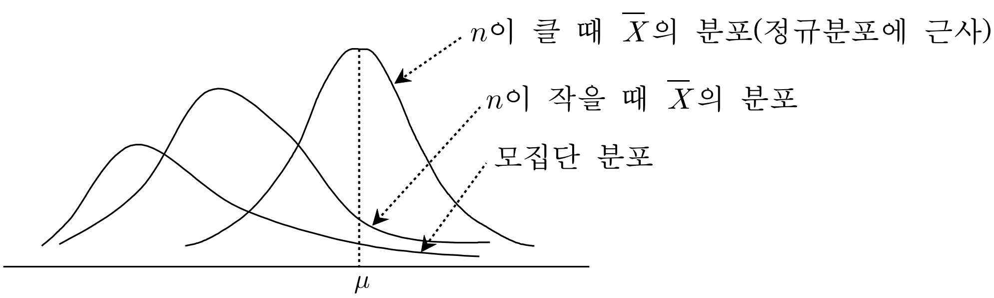
모니터를 생산하는 두 회사가 있다. A사의 모니터는 수명이 평균 6.5년, 표준편차는 0.9년 B사의 모니터는 수명이 평균 6년, 표준편차는 0.8년으로 각각 정규분포를 따른다고 한다. A사에서 36개, B사에서 49개의 확률표본을 랜덤추출했을 때 A사의 표본평균이 B사의 표본평균보다 적어도 1년이상 길 확률을 구하여라.

Thm (36) : 중심극한의 정리(central limit theorem)

평균 μ, 분산 σ^2 인 임의의 모집단으로부터 크기 n 인

표본 $(X_1,\ X_2,\ \cdots\ ,\ X_n)$ 의 평균 $\overline{X}$ 의 분포는 $n \to \infty$ 일 때 $N\left(\mu,\ \dfrac{\sigma^2}{n}\right)$ 에

근사하고 $Z = \dfrac{\overline{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$ 의 분포는 $N(0,\ 1^2)$ 에 근사한다.



예제 219

$\mu = 20$, $\sigma^2 = 16$ 인 무한모집단으로부터 추출한 크기 64 인 표본평균 $\overline{X}$가 19 보다 크고 21 보다 작을 확률을 구하여라.

예제 220

어느 대리점에서 판매원이 고객 한 사람 당 소요되는 시간은 평균 $\mu = 5.4$ 분이고 표준편차 $\sigma = 2.5$ 분이라고 한다. 100 명을 추출하여 조사하였을 때 고객 한 명당 소요된 평균시간이 6 분 이상일 확률을 구하여라.

예제 221

최신 슈퍼 건전지의 수명이 평균 500시간, 표준편차가 35시간인 확률분포를 따른다고 한다. 이 건전지 49개 확률표본의 평균수명이 488시간에서 505시간 사이에 있을 확률을 구하여라.

예제 222

어느 볼펜회사의 매월 볼펜 판매개수는 평균 5650개, 표준편차 700개라고 한다. 임의추출한 볼펜 100개에 대해 그 판매개수를 조사하였을 때 표본평균과 모평균의 차가 200개 이하일 확률을 구하여라.

예제 223

A와 B 두 회사에서 생산되는 형광램프의 평균수명은 각각 1400시간, 1200시간이고 표준편차는 각각 200시간, 100시간이다. 두 회사에서 125개의 형광램프를 추출하여 검사할 때 A회사의 형광램프가 B회사의 형광램프보다 수명이 160시간이상 더 길 확률을 구하여라. (단, $P(0 \le Z \le 2) = 0.4772$)

예제 224

X가 이항분포 $B\left(100,\ \frac{1}{4}\right)$를 따를 때 $P(X > 30)$를 구하여라.

Thm (37): 표본비율의 정규분포

모비율이 p인 어떤 사건이 크기 n인 독립표본 가운데 X개가 나타났다면 표본비율 $\hat{p}=\dfrac{X}{n}$의 분포는 $n\to\infty$일 때 (n이 충분히 클 때)

$\hat{p}=\dfrac{X}{n}$는 근사적으로 $N\left(p,\ \dfrac{pq}{n}\right)$를 따르므로 $Z=\dfrac{\hat{p}-p}{\sqrt{\dfrac{pq}{n}}}$는 근사적으로

$N(0,\ 1^2)$을 따른다. 단) $p+q=1$

$\hat{p}=\dfrac{X}{n}$의 평균 $E(\hat{p})=$

$\hat{p}=\dfrac{X}{n}$의 분산 $\mathrm{Var}(\hat{p})=$

예제 225

어느 도시의 여당 지지율은 전체 시민의 60%라고 한다. 100명의 시민을 랜덤추출하여 조사하였을 때 여당을 지지하는 시민이 50%이하일 확률을 구하여라.

Thm (38) : 자유도 $(n-1)$인 카이제곱분포

정규모집단 $N(\mu,\ \sigma^2)$으로부터 크기 n인 랜덤표본 $(X_1,\ X_2,\ \cdots,\ X_n)$을 추출하였을 때 표본분산을 s^2이라 하면

$\dfrac{(n-1)s^2}{\sigma^2}=\displaystyle\sum_{i=1}^{n}\dfrac{(X_i-\overline{X})^2}{\sigma^2}$ 은 자유도 $(n-1)$인 카이제곱분포를 따른다.

단) $\displaystyle\sum_{i=1}^{n}\dfrac{(X_i-\mu)^2}{\sigma^2}$ 과 $\left(\dfrac{X_i-\mu}{\dfrac{\sigma}{\sqrt{n}}}\right)^2$ 은 서로 독립이라고 알려져 있다.

$\Rightarrow$

<카이제곱분포의 성질>

① $E(X)=v$, $\mathrm{Var}(X)=2v$ ($\because$ v는 자유도)

② χ_1^2이 자유도 v_1인 χ^2분포를 따르고 χ_2^2이 자유도 v_2인 χ^2분포를 따를 때 χ_1^2과 χ_2^2이 서로 독립이면 $\chi_1^2+\chi_2^2$는 자유도 v_1+v_2인 χ^2분포를 따른다.

③ 카이제곱분포는 일반적으로는 비대칭이나 자유도가 10일 경우 거의 대칭이다. 즉, 자유도가 커질수록 정규분포에 근사한다.

④ $P\left[X \ge \chi_\alpha^2(v)\right]=\alpha$로 정의한다.

예제 226

확률변수 X가 자유도 4인 χ^2분포를 따를 때 $P[X \geq x] = 0.05$일 때 x를 구하여라.

예제 227

X가 카이제곱분포를 따를 때 다음을 카이제곱분포표를 이용하여 구하여라.

(1) $\chi^2_{0.90}(60) =$

(2) $\chi^2_{0.01}(14) =$

(3) $\chi^2_{0.05}(9) =$

예제 228

정규모집단 $N(50,\ 2^2)$에서 크기 $n = 16$인 랜덤추출한 표본에서 $P(s^2 > 5)$를 구하여라. (단, s^2은 표본분산)

예제 229

정규모집단 $N(6,\ 2^2)$에서 크기 $n = 10$인 랜덤추출한 표본분산 s^2이 5보다 클 확률을 구하여라.

예제 230

X가 $\chi^2(10)$을 따를 때 $P(3.25 \leq X \leq 20.5)$를 구하여라.

예제 231

X가 $\chi^2(10)$를 따를 때 $P(X > a) = 0.05$일 때 a를 구하여라.

예제 232

X가 $\chi^2(5)$분포를 따를 때 $P\left(X < \chi_\alpha^2\right) = 0.95$가 되는 χ_α^2를 구하여라.

예제 233

X가 $\chi^2(20)$을 따를 때 $P(9.591 < X < 34.170)$를 구하여라.

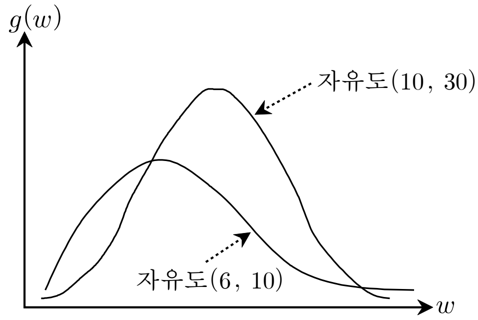
Def (41):F-분포

서로 독립인 두 확률변수 U와 V가 각각 자유도 r_1, r_2인 카이제곱분포를 따를 때 $W=\dfrac{U/r_1}{V/r_2}$는 자유도 $(r_1,\ r_2)$인 F 분포를 따른다고 한다.

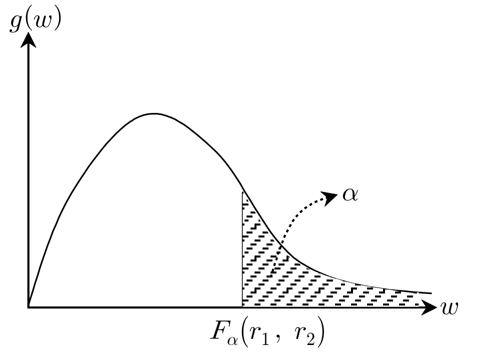
W의 확률밀도함수

$$g(w)=\frac{\Gamma\left(\frac{r_1+r_2}{2}\right)\cdot\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}}}{\Gamma\left(\frac{r_1}{2}\right)\cdot\Gamma\left(\frac{r_2}{2}\right)}\cdot w^{\frac{r_1}{2}-1}\cdot\left(1+\frac{r_1}{r_2}w\right)^{-\frac{r_1+r_2}{2}},\ w>0$$ 이다.

※ 자유도 r_1, r_2의 변화에 따른 F 분포



※ $P[W\geq F_\alpha(r_1,\ r_2)]=\alpha$로 정의한다.



예제 234

확률변수 X가 자유도 $(5,\ 10)$인 F분포를 따를 때 $P[W\geq a]=0.05$를 만족하는 a를 구하여라.

Thm (39)

$F_{1-\alpha}(r_1,\ r_2)=\dfrac{1}{F_\alpha(r_2,\ r_1)}$ 이다.

$\Rightarrow$

예제 235

확률변수 X가 자유도 $(5,\ 10)$인 F분포를 따르 때 $P(X \leq a) = 0.01$을 만족하는 a를 구하여라. (단, $F_{0.99}(10,\ 5) = 10.1$이다.)

예제 236

확률변수 $X \sim F(6,\ 10)$일 때 $P(X \geq a) = 0.95$를 만족하는 a를 구하여라.
(단, $F_{0.05}(10,\ 6) = 4.06$)

Thm (40): 자유도 $(n_1 - 1,\ n_2 - 1)$인 F분포

모분산이 각각 $\sigma_1^2,\ \sigma_2^2$인 정규모집단에서 서로 독립적으로 추출된 크기 $n_1,\ n_2$인 표본의 분산을 $s_1^2,\ s_2^2$이라 할 때

$$F = \frac{\dfrac{s_1^2}{\sigma_1^2}}{\dfrac{s_2^2}{\sigma_2^2}} = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2}$$ 은 자유도 $(n_1 - 1,\ n_2 - 1)$인 F분포를 따른다.

$\Rightarrow$

예제 237

모분산이 같은 정규분포에서 표본의 크기가 각각 $n_1=8$, $n_2=12$ 인 서로 독립인 표본을 임의추출하였을 때 각각의 표본분산을 s_1^2, s_2^2 이라고 하면 확률 $P\left(\frac{s_1^2}{s_2^2}>4.89\right)$를 구하여라.

예제 238

확률변수 X가 자유도 $(7,\ 8)$ 인 F분포를 따를 때 $P(X>a)=0.05$, $P(X\leq b)=0.99$ 인 a와 b를 구하여라.

예제 239

정규분포 $N(50,\ 2^2)$ 에서 각각 표본크기 $n_1=9$, $n_2=16$ 인 서로 독립인 표본을 랜덤추출하였을 때 각각의 표본분산 s_1^2, s_2^2 이라 두면 확률 $P\left(\frac{s_1^2}{s_2^2}>4\right)$를 구하여라.

Def (42): $t-$**분포**

Z는 표준정규분포를 따르고 V는 자유도 r인 카이제곱분포를 따를 때 서로 독립인 Z와 V에 대하여 새로운 확률변수

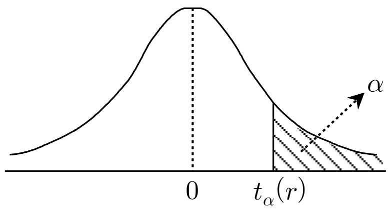
$T=\dfrac{Z}{\sqrt{\dfrac{V}{r}}}$ 는 자유도 r인 $t-$ 분포를 따른다고 한다.

T의 확률밀도함수 $h(t)=\dfrac{\Gamma\left(\dfrac{r+1}{2}\right)}{\Gamma\left(\dfrac{r}{2}\right)\sqrt{\pi r}}\left(1+\dfrac{t^2}{r}\right)^{-\frac{r+1}{2}}$, $-\infty<t<\infty$

※ t분포는 원점 중심 좌우대칭이다.

$\Rightarrow$

※ $P\left[T>t_{\alpha}(r)\right]=\alpha$로 정의한다. ($t$분포는 원점 중심 좌우대칭이다.)



예제 240

다음을 계산하여라.

(1) T는 자유도 19인 t분포일 때 $P(T\geq t)=0.025$를 만족하는 t를 구하여라.

(2) T는 자유도 7인 t분포일 때 $P(T\leq -1.415)$를 구하여라.

Thm (41)

평균이 μ이고 분산이 σ^2인 정규분포에서 랜덤추출한 표본 $X_1,\ X_2,\ \cdots,\ X_n$에 대하여

확률변수 $T = \dfrac{\overline{X} - \mu}{\dfrac{s}{\sqrt{n}}}$는 자유도 $n-1$인 t분포를 따른다.

단) $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

$\Rightarrow$

※ t분포의 성질 (자유도 n)

(1) $E(X) = 0,\ \mathrm{Var}(X) = \dfrac{n}{n-2}\ (n > 2)$

(2) t분포는 0을 중심으로 좌우 대칭으로 $n \to \infty$일 때 표준정규분포 $N(0,\ 1)$에 가까워진다.

(3) 표본의 크기 n이 작을 때 보통 $n < 30$일 때 t분포를 사용한다.

예제 241

정규분포에서 15개의 표본을 추출한 표본 X_1, X_2, $\cdots$, X_{15}에 대하여 확률변수 $T = \dfrac{\overline{X} - \mu}{\dfrac{\sigma}{\sqrt{n}}}$ 라 하자. 이 때 $P(k < T < -1.761) = 0.045$을 만족하는 k값을 구하여라.

예제 242

T가 자유도 n인 t분포를 따를 때 T^2은 자유도 $(1,\ n)$인 F분포를 따름을 증명하여라.

〈심화 특강 ②〉

(1) F-분포의 확률밀도함수

서로 독립인 두 확률변수 U과 V가 각각 자유도 r_1, r_2인 카이제곱분포를 따를 때 확률변수 $W=\dfrac{U/r_1}{V/r_2}$는 자유도 (r_1, r_2)인 F분포를 따른다. 이 때, W의 확률밀도함수

$$g(w)=\frac{\Gamma\left(\frac{r_1+r_2}{2}\right)\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}}}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)}\cdot w^{\frac{r_1}{2}-1}\left(1+\frac{r_1}{r_2}w\right)^{-\frac{r_1+r_2}{2}},\quad w>0$$ 임을 증명하여라.

(증명)

서로 독립인 두 확률변수 U와 V는 자유도가 r_1, r_2이고 χ^2분포를 따르므로
U와 V의 결합확률밀도함수는

$$h(u,\ v)=h(u)h(v)=\frac{1}{\Gamma\left(\frac{r_1}{2}\right)2^{\frac{r_1}{2}}}\cdot u^{\frac{r_1}{2}-1}e^{-\frac{u}{2}}\times\frac{1}{\Gamma\left(\frac{r_2}{2}\right)2^{\frac{r_2}{2}}}\cdot v^{\frac{r_2}{2}-1}\cdot e^{-\frac{v}{2}}$$

$$=\frac{1}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)2^{\frac{r_1}{2}+\frac{r_2}{2}}}\cdot u^{\frac{r_1}{2}-1}v^{\frac{r_2}{2}-1}e^{-\frac{u+v}{2}},\quad 0<u,\ v<\infty$$ 이다.

이제, 새로운 확률변수 $W=\dfrac{U/r_1}{V/r_2}$라 두고 W의 주변밀도함수 $g_1(w)$를 구해보자.

$w=\dfrac{u/r_1}{v/r_2}$, $z=v$는 $(u,\ v)$를 $(w,\ z)$로 일대일 대응 변환시킨다.

$v=z$, $u=\left(\dfrac{r_1}{r_2}\right)zw$ 이므로 $J=\begin{vmatrix}\frac{\partial u}{\partial w} & \frac{\partial u}{\partial z}\\ \frac{\partial v}{\partial w} & \frac{\partial v}{\partial z}\end{vmatrix}=\begin{vmatrix}\frac{r_1}{r_2}z & \frac{r_1}{r_2}w\\ 0 & 1\end{vmatrix}=\dfrac{r_1}{r_2}z$ $(\because\ v>0,\ z>0)$

$|J|=\dfrac{r_1}{r_2}z$ 이다.

$$g(w,\ z)=h\left(\left(\frac{r_1}{r_2}\right)zw,\ z\right)|J|=\frac{1}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)2^{\frac{r_1+r_2}{2}}}\cdot\left(\frac{r_1zw}{r_2}\right)^{\frac{r_1}{2}-1}\cdot(z)^{\frac{r_2}{2}-1}\cdot e^{-\frac{z}{2}\left(\frac{r_1w}{r_2}+1\right)}\times\frac{r_1}{r_2}z\ ,$$

$$0<w,\ z<\infty$$

따라서, $g_1(w)=\displaystyle\int_0^\infty g(w,\ z)\,dz$ 을 적분하면

※여기서, $\dfrac{z}{2}\left(\dfrac{r_1}{r_2}w+1\right)=y$라 두면 $\Rightarrow z=\dfrac{2y}{\frac{r_1}{r_2}w+1}$ 대입 $\Rightarrow dz=\dfrac{2}{\frac{r_1}{r_2}w+1}dy$

$$\therefore\ g_1(w)=\frac{\Gamma\left(\frac{r_1+r_2}{2}\right)\left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}}}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)}\cdot w^{\frac{r_1}{2}-1}\left(1+\frac{r_1}{r_2}w\right)^{-\frac{r_1+r_2}{2}},\ 0<w<\infty$$ 로 정리된다.

(2) F-분포의 평균

W가 자유도 $(n,\ m)$인 F분포를 따를 때 $E(W)=\dfrac{m}{m-2}$, $m>2$임을 증명하여라.

또, $W\sim F(5,\ 8)$일 때 $E(W)$를 구하여라.

(풀이)

$\Rightarrow$ W가 자유도 $(n,\ m)$인 F분포이므로 $W=\dfrac{\frac{U}{n}}{\frac{V}{m}}$라 두면 U와 V는 서로 독립이고 각각 자유도 $n,\ m$인 카이제곱분포를 따른다.

(참고) W가 자유도 $(n,\ m)$인 F분포를 따를 때 $\mathrm{Var}(W)=\dfrac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$이다.

(3) t-분포의 확률밀도함수

Z는 $N(0,\ 1)$을 따르는 확률변수이고 V는 $\chi^2(r)$를 따르는 확률변수일 때 서로 독립인 Z와 V에 대하여 확률변수 $T=\dfrac{Z}{\sqrt{\dfrac{V}{r}}}$ 는 자유도 r인 t분포를 따른다. 이 때

T의 확률밀도함수 $g(t)=\dfrac{\Gamma\left(\dfrac{r+1}{2}\right)}{\Gamma\left(\dfrac{r}{2}\right)\sqrt{\pi r}}\left(1+\dfrac{t^2}{r}\right)^{-\frac{r+1}{2}}$, $-\infty<t<\infty$

(증명)

Z와 V는 서로 독립이므로 Z와 V의 결합확률밀도함수

$$h(z,\ v)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\cdot\frac{1}{\Gamma\left(\frac{r}{2}\right)2^{\frac{r}{2}}}\cdot v^{\frac{r}{2}-1}\cdot e^{-\frac{v}{2}},\ -\infty<z<\infty,\ 0<v<\infty$$

$T=\dfrac{Z}{\sqrt{\dfrac{V}{r}}}$ 라 두면 $t=\dfrac{z}{\sqrt{\dfrac{v}{r}}}$, $u=v$는 $(z,\ v)$는 $(t,\ u)$로 일대일 대응 변환을 정의한다.

$z=t\sqrt{\dfrac{u}{r}}$, $v=u$에서 $J=\begin{vmatrix}\dfrac{\partial z}{\partial t} & \dfrac{\partial z}{\partial u}\\ \dfrac{\partial v}{\partial t} & \dfrac{\partial v}{\partial u}\end{vmatrix}=\begin{vmatrix}\sqrt{\dfrac{u}{r}} & t\dfrac{\frac{1}{r}}{2\sqrt{\frac{u}{r}}}\\ 0 & 1\end{vmatrix}\Rightarrow |J|=\sqrt{\dfrac{u}{r}}$ 이다.

$$g(t,\ u)=h\left(t\sqrt{\frac{u}{r}},\ u\right)|J|=\frac{1}{\sqrt{2\pi}\,\Gamma\left(\frac{r}{2}\right)2^{\frac{r}{2}}}\cdot e^{-\frac{u}{2}\left(\frac{t^2}{r}+1\right)}\cdot u^{\frac{r}{2}-1}\cdot\sqrt{\frac{u}{r}},\ -\infty<t<\infty,\ 0<u<\infty$$

※ $e^{-\frac{z^2}{2}}\times e^{-\frac{v}{2}}=e^{-\frac{1}{2}(z^2+v)}$ $\left(\leftarrow z=t\sqrt{\dfrac{u}{r}},\ v=u\right)\Rightarrow e^{-\frac{1}{2}\left(t^2\frac{u}{r}+u\right)}=e^{-\frac{1}{2}u\left(\frac{t^2}{r}+1\right)}$

$g_1(t)=\displaystyle\int_0^\infty g(t,\ u)du$ 여기서 $\dfrac{u}{2}\left(\dfrac{t^2}{r}+1\right)=w$라 치환적분하면

$\therefore\ g_1(t)=\dfrac{\Gamma\left(\dfrac{r+1}{2}\right)}{\Gamma\left(\dfrac{r}{2}\right)\sqrt{\pi r}}\left(1+\dfrac{t^2}{r}\right)^{-\frac{r+1}{2}}$, $-\infty<t<\infty$ 이다.

제 8 장　추정 이론

Def (43): 추정량과 추정값

표본의 데이터로부터 미지의 모수 θ를 추정하기 위하여 이용되는 통계량 $\hat{\theta}(X_1,\ X_2,\ \cdots,\ X_n)$을 θ의 **추정량**이라고 하고 $\hat{\theta}$의 측정값 $\hat{\theta}(x_1,\ x_2,\ \cdots,\ x_n)$을 θ의 **추정값**이라고 한다.

$\Rightarrow$

예제 243

$X_1,\ X_2,\ \cdots,\ X_n$을 $N(\mu,\ \sigma^2)$으로부터 얻은 랜덤표본이라 할 때 모평균 μ와 모분산 σ^2을 추정하고자 한다. 각각의 추정량과 추정값을 구하여라.

예제 244

모평균 μ를 추정하기 위해서 모집단에서 추출한 확률표본 $X_1,\ X_2,\ X_3$에 대해
다음과 같은 두 가지 추정량을 생각했다.

$$\widehat{\mu_1} = \frac{X_1+X_2+X_3}{3},\qquad \widehat{\mu_2} = \frac{X_1+3X_2+5X_3}{9}$$

이 때 각각의 관측값 $x_1=3,\ x_2=4,\ x_3=5$를 얻었을 때 각각의 추정량에 대한 추정값을 구하여라.

Thm (42) : 불편추정량(Unbiased estimator)

$T(X)$를 $g(\theta)$의 추정량이라 할 때 $E[T(X)]-g(\theta)$를 $T(X)$의 **편향(bias)**이라 하고 $E[T(X)]=g(\theta)$ 즉, 편향$=0$이면 $T(X)$를 $g(\theta)$의 **불편추정량**이라고 한다.

$\Rightarrow$

예제 245

표본분산 s^2은 모분산 σ^2의 불편추정량임을 보여라.

예제 246

평균 μ와 분산 σ^2을 갖는 정규모집단에서 추출한 크기 n의 확률표본 Y_1, Y_2, $\cdots$, Y_n이라 할 때 표본분산 s^2이 모분산 σ^2의 불편추정량임을 보여라.

예제 247

평균 μ와 분산 σ^2을 갖는 정규분포에서 추출한 확률변수 X_1, X_2, X_3에 대하여 두 가지 추정량 $\widehat{\mu_1}=\dfrac{X_1+X_2+X_3}{3}$, $\widehat{\mu_2}=\dfrac{X_1+3X_2+5X_3}{9}$이 모평균 μ의 불편추정량임을 보여라.

Thm (43) : 유효추정량(efficient estimator)

$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$ 이면 $\hat{\theta}_1$ 은 $\hat{\theta}_2$ 보다 유효하다고 한다.

즉, 추정량의 분산이 작을수록 더 유효하다.

예제 248

평균 μ 와 분산 σ^2 을 갖는 정규분포에서 추출한 확률변수 X_1, X_2, X_3 에 대하여 두 추정량 $\hat{\mu_1} = \dfrac{X_1 + X_2 + X_3}{3}$, $\hat{\mu_2} = \dfrac{X_1 + 3X_2 + 5X_3}{9}$ 은 각각 μ 의 불편추정량이다. 이 두 추정량 중 어느 것이 더 유효한 추정량인가?

Def (44) : 피셔의 정보(Fisher's information)

$$I(\theta)=\int_{-\infty}^{\infty}\frac{\partial \log f(x\,;\,\theta)}{\partial\theta}\cdot\frac{\partial \log f(x\,;\,\theta)}{\partial\theta}\cdot f(x\,;\,\theta)\ dx$$

$$=E\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right]$$ 을 **피셔의 정보**라 한다.

$I(\theta)$는 결국 확률변수 $\dfrac{\partial \log f(X;\theta)}{\partial\theta}$의 분산이 된다.

$\Rightarrow$

예제 249

정규확률변수에 대한 정보 $X\sim N(\mu,\ \sigma_0^2)$일 때 $I(\mu)$를 구하여라.

예제 250

베르누이 확률변수에 대한 정보 X가 성공확률이 p인 베르누이 확률변수라 할 때 $I(p)$를 구하여라.

Thm (44) : 라오-크래머 하한(정보부등식)

확률밀도함수 $f(x\,;\theta)$와 통계량 $T(X)$에 대해 ①과 ②를 만족한다.

① $A=\{x\ ;\ f(x\,;\theta)>0\}$는 모수 θ에 의존하지 않고

모든 $x\in A$, $\theta\in\Omega$에 대하여 $\dfrac{\partial}{\partial\theta}\log f(x\,;\theta)$는 유한하게 존재한다.

② $g(\theta)=\displaystyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}t(x_1,\ x_2,\ \cdots,\ x_n)f(x_1\,;\theta)\ \cdots\ f(x_n\,;\theta)\ dx_1\ \cdots\ dx_n$

이 때, $\mathrm{Var}(T(X))\geq\dfrac{[g'(\theta)]^2}{nI(\theta)}$ 이 성립한다.

특히, $T(X)$가 θ의 불편 추정량이면 $\mathrm{Var}(\hat{\theta})\geq\dfrac{1}{nI(\theta)}$ 이다.

※ 최소 분산 불편추정량 (Minimum Variance Unbiased Estimator)

⇒ 확률변수 $X_1,\ X_2,\ \cdots,\ X_n$의 결합확률밀도함수가 $f(x_1,\ x_2,\ \cdots,\ x_n)$이라 할 때 모수의 함수 $g(\theta)$의 추정 $T^*(X)$가 다음 두 조건을 모두 만족하면 최소불편추정량이라 한다.

① $T^*(X)$는 $g(\theta)$의 불편추정량이다. 즉, $E(T(X))=g(\theta)$이다.

② $g(\theta)$의 임의의 다른 불편추정량 $T(X)$에 대해 $\mathrm{Var}\,T^*(X)\leq\mathrm{Var}\,T(X)$

예제 251

X_1, X_2, $\cdots$, X_n을 성공확률이 p인 베르누이 확률분포에서 구한 랜덤표본이라 할 때 p의 최소분산불편추정량을 구하여라.

예제 252

X_1, X_2, $\cdots$, X_n을 $N(\mu,\ \sigma^2)$으로부터 추출한 랜덤 표본이라할 때 μ의 최소분산불편추정량을 구하여라.

예제 253

X_1, X_2, $\cdots$, X_n을 모수가 λ인 포아송 분포에서 추출한 랜덤표본이라 할 때 λ의 최소분산불편추정량을 구하여라.

예제 254

Thm (44) 에서 $\mathrm{Var}(T(X)) \geq \dfrac{[g'(\theta)]^2}{nI(\theta)}$ 임을 증명하여라. (증명)

$$g'(\theta) = \frac{\partial}{\partial\theta} g(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} t(x_1, x_2, \cdots, x_n) \frac{\partial}{\partial\theta}\left[\prod_{i=1}^{n} f(x_i;\theta)\right] dx_1 \cdots dx_n$$

$$- g(\theta)\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^{n} f(x_i;\theta)\, dx_1 \cdots dx_n$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [t(x_1, \cdots, x_n) - g(\theta)] \frac{\partial}{\partial\theta}\left[\prod_{i=1}^{n} f(x_i;\theta)\right] dx_1 \cdots dx_n$$

※ $\dfrac{\partial}{\partial\theta}\left[\prod_{i=1}^{n} f(x_i;\theta)\right] = \left[\sum_{i=1}^{n}\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right] \cdot \prod_{i=1}^{n} f(x_i;\theta)$

그 예로 $\dfrac{\partial}{\partial\theta}\left[\prod_{i=1}^{3} f(x_i;\theta)\right] = \left[\sum_{i=1}^{3}\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right] \cdot \prod_{i=1}^{3} f(x_i;\theta)$ 임을 알아보자.

좌변 $= \dfrac{\partial}{\partial\theta}\left[\prod_{i=1}^{3} f(x_i;\theta)\right] = \dfrac{\partial}{\partial\theta}[f(x_1;\theta)f(x_2;\theta)f(x_3;\theta)]$

$= \dfrac{\partial}{\partial\theta}f(x_1;\theta)\, f(x_2;\theta)f(x_3;\theta) + f(x_1;\theta)\dfrac{\partial}{\partial\theta}f(x_2;\theta)f(x_3;\theta) + f(x_1;\theta)\, f(x_2;\theta)\dfrac{\partial}{\partial\theta}f(x_3;\theta)$

우변 $= \left[\sum_{i=1}^{3}\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right]\prod_{i=1}^{3} f(x_i;\theta) = \left[\sum_{i=1}^{3}\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right] \times f(x_1;\theta)f(x_2;\theta)f(x_3;\theta)$

$= \left[\dfrac{\frac{\partial}{\partial\theta}f(x_1;\theta)}{f(x_1;\theta)} + \dfrac{\frac{\partial}{\partial\theta}f(x_2;\theta)}{f(x_2;\theta)} + \dfrac{\frac{\partial}{\partial\theta}f(x_3;\theta)}{f(x_3;\theta)}\right] \times f(x_1;\theta)f(x_2;\theta)f(x_3;\theta)$

$= \dfrac{\partial}{\partial\theta}f(x_1;\theta)\, f(x_2;\theta)f(x_3;\theta) + f(x_1;\theta)\dfrac{\partial}{\partial\theta}f(x_2;\theta)f(x_3;\theta) + f(x_1;\theta)\, f(x_2;\theta)\dfrac{\partial}{\partial\theta}f(x_3;\theta)$

따라서, 좌변=우변이 된다. 확장하여 $\dfrac{\partial}{\partial\theta}\left[\prod_{i=1}^{n} f(x_i;\theta)\right] = \left[\sum_{i=1}^{n}\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right] \cdot \prod_{i=1}^{n} f(x_i;\theta)$가 된다.

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [t(x_1, \cdots, x_n) - g(\theta)] \left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i;\theta)\right]\prod_{i=1}^{n} f(x_i;\theta)\, dx_1 \cdots dx_n$$

$$= E\left[\,[t(x_1, \cdots, x_n) - g(\theta)]\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i;\theta)\right]\right]$$

따라서, $g'(\theta) = E\left[\,[t(x_1, \cdots, x_n) - g(\theta)]\left[\sum_{i=1}^{n}\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right]\right]$이다.

코시-슈바르츠 정리(즉, $[E(XY)^2 \leq E(X^2)E(Y^2)]$)에 의해

$$[g'(\theta)]^2 \leq E\{[t(x_1, \cdots, x_n) - g(\theta)]^2\} E\left\{\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i;\theta)\right]^2\right\}$$

$$\mathrm{Var}(T(X)) \geq \frac{[g'(\theta)]^2}{E\left\{\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i;\theta)\right]^2\right\}} = \frac{[g'(\theta)]^2}{nI(\theta)}$$

※ $E\left\{\left[\sum_{i=1}^{n}\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right]^2\right\} = E(\square^2) = \mathrm{Var}(\square) + \{E(\square)\}^2 = nI(\theta)$

($\because$) Def (44) 에서 $E\left[\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right] = 0$, $\mathrm{Var}\left[\dfrac{\partial}{\partial\theta}\log f(x_i;\theta)\right] = I(\theta)$ 이므로

$E(\square) = E(\triangle_1 + \triangle_2 + \cdots + \triangle_n) = E(\triangle_1) + \cdots + E(\triangle_n) = 0$

$\mathrm{Var}(\square) = \mathrm{Var}(\triangle_1 + \triangle_2 + \cdots + \triangle_n) = \mathrm{Var}(\triangle_1) + \cdots + \mathrm{Var}(\triangle_n) = nI(\theta)$

Def (45) : 충분통계량(sufficient statistic)

확률표본 $X_1,\ X_2,\ \cdots\ ,\ X_n$ 의 결합확률밀도함수가 $f(x_1,\ x_2,\ \cdots\ ,\ x_n\,;\theta)$ 일 때 통계량 $Y_1 = u(X_1,\ X_2,\ \cdots\ ,\ X_n)$ 이

$$\frac{f(x_1,\ x_2,\ \cdots\ ,\ x_n\,;\theta)}{f_{Y_1}[u(x_1,\ x_2,\ \cdots\ ,\ x_n)\,;\theta]} = h(x_1,\ x_2,\ \cdots\ ,\ x_n) \quad \text{일 때}$$

θ 에 대한 **충분통계량**이라 한다.

$\Rightarrow$ 모수 θ 에 대한 정보가 모두 충분통계량 $Y = u(X_1,\ X_2,\ \cdots\ ,\ X_n)$ 에 포함되어 있다는 말이다.

$\Rightarrow$ $f(x_1,\ x_2,\ \cdots\ ,\ x_n\,;\theta) = f(x_1\,;\theta)\ \cdots\cdots\ f(x_n\,;\theta)$ 인 경우로 한정.

예제 255

$X_1,\ X_2,\ \cdots\ ,\ X_n$ 을 베르누이(p) 분포에서 추출한 확률표본이라 할 때
통계량 $Y_1 = X_1 + X_2 + \ \cdots\ + X_n$ 은 충분통계량임을 보여라.

예제 256

X_1, X_2, $\cdots$, X_n을 $\alpha=2$, $\beta=\theta$ (>0)인 감마분포에서 얻은 확률표본이라 할 때 $Y_1=\sum_{i=1}^{n} X_i$는 θ에 대한 충분통계량임을 보여라.

Def (46) : 네이만의 인수분해정리에 의한 충분통계량

$X_1,\ X_2,\ \cdots,\ X_n$을 모수 θ인 확률분포에서 얻은 랜덤표본이라 할 때

통계량 $Y_1 = u_1(X_1,\ X_2,\ \cdots,\ X_n)$은

$f(x_1\,;\theta)f(x_2\,;\theta)\ \cdots\ f(x_n\,;\theta) = k_1[u_1(x_1,\ x_2,\ \cdots,\ x_n)\,;\theta]\,k_2(x_1,\ x_2,\ \cdots,\ x_n)$

을 만족하는 2개의 음이 아닌 함수 k_1과 k_2를 구할 수 있을 때

만약, $k_2(x_1,\ x_2,\ \cdots,\ x_n)$이 θ에 종속되지 않는다면

θ에 대한 충분통계량이다.

예제 257

$X_1,\ X_2,\ \cdots,\ X_n$을 모수 λ을 갖는 포아송분포에서 얻은 확률표본이라 할 때

통계량 $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$는 λ에 대한 충분통계량임을 보여라.

예제 258

$X_1,\ X_2,\ \cdots\ ,\ X_n$을 $N(\mu,\ 1),\ -\infty < \mu < \infty$로부터 얻은 확률표본이라 할 때 $\overline{X}$는 μ에 대한 충분통계량임을 보여라.

예제 259

$X_1,\ X_2,\ \cdots\ ,\ X_n$이 베르누이 (p)분포에서 얻은 확률표본일 때 $Y_1 = \sum_{i=1}^{n} X_i$는 p에 대한 충분통계량임을 보여라.

〈심화 특강 ③〉

〈지수형태 확률밀도함수에서의 충분통계량〉

$X_1,\ X_2,\ \cdots,\ X_n$ 이 다음과 같은 지수형태의 확률밀도함수 $f(x\,;\theta)=\exp[k(x)p(\theta)+s(x)+q(\theta)]$ 를 갖는 분포로부터 얻은 확률표본이라 할 때 집합 $\{x \mid f(x\,;\theta)>0\}$ 는 모수 θ 에 의존하지 않는다면 통계량 $\sum_{i=1}^{n} k(X_i)$ 는 θ 에 대한 충분통계량이다.

$\Rightarrow$

ex 1)

$X_1,\ X_2,\ \cdots,\ X_n$ 이 모수 θ 를 갖는 지수분포에서 얻은 확률분포일 때 $\overline{X}=\dfrac{1}{n}\sum_{i=1}^{n} x_i$ 는 θ 에 대한 충분통계량임을 보여라.

ex 2)

X_1, X_2, $\cdots$, X_n이 정규분포 $N(\mu,\ \sigma^2)$에서 얻은 확률표본일 때 $\overline{X}$와 s^2은 μ와 σ^2에 대한 결합충분통계량임을 보여라.

〈라오-블랙웰 정리〉

Y_1이 모수 θ의 충분통계량이고 Y_2가 θ의 불편추정량일 때
$\delta(Y_1)=E(Y_2/Y_1)$라 두면 $\delta(Y_1)$도 역시 θ의 불편추정량이고
모든 θ에 대하여
$\mathrm{Var}(\delta(Y_1))=E\left[\{\delta(Y_1)-\theta\}^2\right]\leq E\left[(Y_2-\theta)^2\right]=\mathrm{Var}(Y_2)$ 이다.

$\Rightarrow$

① 어떤 불편 추정량 Y_2가 존재하면 충분통계량 Y_1에 대한 조건부 기댓값 $E(Y_2/Y_1)$ 역시 불편 추정량이고 Y_2보다 작거나 같은 분산을 가진다.
따라서, 불편추정량에서 작은 분산을 가지는 추정량을 찾을 때 충분통계량의 함수꼴에 국한시킬 수 있다.
② 특히, 확률밀도함수가 지수형태일 때 만약 불편추정량이 존재한다면 그것은 충분통계량의 유일한 하나의 불편함수가 존재한다는 것이다.
③ 모수 θ에 대한 충분통계량 Y_1과 관련하여 확률밀도함수가 지수형태일 때는 만약 다른 통계량 Z가 θ와 무관한 분포를 가진다면 Y와 Z는 독립인 것이다.

(증명)
이중기댓값정리에 의해 $E(\delta(Y_1))=E[E(Y_2/Y_1)]=E(Y_2)=\theta$이므로 $\delta(Y_1)$도 θ의 불편추정량이다.
$\mathrm{Var}(Y_2)=E\left[(Y_2-E(Y_2))^2\right]=E\left[(Y_2-\theta)^2\right]$
$=E\left[(Y_2-\delta(Y_1)+\delta(Y_1)-\theta)^2\right]$
$=E\left[(Y_2-\delta(Y_1))^2\right]+E\left[(\delta(Y_1)-\theta)^2\right]+2E[(Y_2-\delta(Y_1))(\delta(Y_1)-\theta)]$ …㈀

여기서, $E[(Y_2-\delta(Y_1))(\delta(Y_1)-\theta)]=E\{E[(Y_2-\delta(Y_1))(\delta(Y_1)-\theta)/Y_1]\}=0$
($\because$) $E[(Y_2-\delta(Y_1))(\delta(Y_1)-\theta)/Y_1]\}$
$=[\delta(Y_1)-\theta]\,E[(Y_2-\delta(Y_1))/Y_1]$
$=[\delta(Y_1)-\theta]\,[E(Y_2/Y_1)-E(\delta(Y_1)/Y_1)]$
$=[\delta(Y_1)-\theta]\,[E(Y_2/Y_1)-E(Y_2/Y_1)]=0$

따라서, ㈀식$=E\left[(Y_2-\delta(Y_1))^2\right]+\mathrm{Var}(\delta(Y_1))$
$\therefore$ $\mathrm{Var}(Y_2)\geq\mathrm{Var}(\delta(Y_1))$

ex 3)

X_1, X_2, $\cdots$, X_n 을 $N(\mu,\ \sigma^2)$ 에서 얻은 확률표본이라 할 때 μ 의 최소분산불편추정량을 구하여라.

Def (47) : 신뢰구간(confidence interval)

랜덤 표본 $X_1,\ X_2,\ \cdots\ ,\ X_n$의 모확률밀도함수가 $f(x\,;\theta),\ \theta\in\Omega$ 라고 하자. 확률구간 $\left[\widehat{\theta_L},\ \widehat{\theta_U}\right]$ 에 대하여

$P\left(\widehat{\theta_L}\le\theta\le\widehat{\theta_U}\right)=1-\alpha$ (단, $0<\alpha<1$)를 만족시키면 이를 모수 θ의 $100\cdot(1-\alpha)\%$ **신뢰구간**이라고 부른다.

여기서, $\widehat{\theta_L}$을 신뢰구간의 하한, $\widehat{\theta_U}$을 신뢰구간의 상한이라고 부른다.

또 확률 $1-\alpha$를 **신뢰도** or **신뢰계수**라 부른다.

$\Rightarrow$ $P\left(\widehat{\theta_L}\le\theta\right)=1-\alpha$일 때 구간 $\left[\widehat{\theta_L},\ \infty\right)$를 θ의 단측하한 신뢰구간

$P\left(\theta\le\widehat{\theta_U}\right)=1-\alpha$일 때 구간 $\left(-\infty,\ \widehat{\theta_U}\right]$를 θ의 단측상한 신뢰구간

예제 260

$X_1,\ X_2,\ \cdots\ ,\ X_n$을 $N(\mu,\ 25)$로부터 얻은 확률표본이라 할 때

통계량 $\widehat{\theta_L}=\overline{X}-\dfrac{10}{\sqrt{n}}$, $\widehat{\theta_U}=\overline{X}+\dfrac{10}{\sqrt{n}}$이라 할 때 $P\left(\widehat{\theta_L}\le\theta\le\widehat{\theta_U}\right)=1-\alpha$를 만족하는 모평균 μ에 대한 신뢰도를 구하여라.

Thm (45) : 모평균 μ의 구간 추정① (모분산 σ^2을 알고있는 경우)

모분산 σ^2이 알려진 모집단에서 얻은 크기 n인 확률표본의 평균을 $\overline{X}$라 할 때 μ의 $100(1-\alpha)\%$ 신뢰구간은

$$\overline{X}-z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\leq\mu\leq\overline{X}+z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$$ 이다.

여기서, $z_{\frac{\alpha}{2}}$는 오른쪽 면적이 $\frac{\alpha}{2}$인 z값이다. 또, 모집단은 정규모집단이다.

$\Rightarrow$ μ의 추정값으로 $\overline{X}$가 사용될 때

(1) 오차가 $Z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$을 초과하지 않는다고 $100(1-\alpha)\%$ 확신할 수 있다.

(2) 표본의 크기 $n=\left(\frac{z_{\frac{\alpha}{2}}\cdot\sigma}{e}\right)^2$일 때 오차는 특정한 값 e를 초과하지 않는다고 $100(1-\alpha)\%$ 확신할 수 있다.

예제 261

어떤 전구공장에서 생산되는 전구의 수명시간은 표준편차가 36시간으로 알려져 있다. 임의로 27개를 추출하여 전구의 수명을 재어보니 1478시간이었다. 이 공장에서 생산되는 전구의 평균수명시간에 대한 95%의 신뢰구간을 구하여라. (단, 수명시간은 정규분포를 따른다.)

예제 262

어느 전구 제조회사에서 생산되는 전구의 수명을 X라 하자.

X는 $N(\mu,\ 120^2)$을 따른다고 알려져 있다. 100개의 확률표본을 얻었을 때 $\overline{X}=2000$이었다. μ의 95% 신뢰구간을 구하여라.

예제 263

어떤 농지의 중금속 오염농도 X는 $N(\mu,\ 0.3^2)$을 따른다고 알려져 있다.

36곳의 확률표본을 얻었을 때 $\overline{X}=2.6$이라면

(1) 95% 신뢰구간과 99%신뢰구간을 구하여라.

(2) 95%확신으로 μ의 추정값이 0.05이하의 오차를 가질 때 표본의 크기는 최소 얼마인가?

Thm (46) : μ의 단측 신뢰한계(σ^2이 알려져 있는 경우)

분산 σ^2인 정규모집단에서 얻은 크기 n인 표본의 평균을 $\overline{X}$라 할 때 μ에 대한 $100(1-\alpha)\%$ 단측신뢰한계는

신뢰상한 $\Rightarrow$ $\overline{X}+z_\alpha \cdot \dfrac{\sigma}{\sqrt{n}}$, 신뢰하한 $\Rightarrow$ $\overline{X}-z_\alpha \cdot \dfrac{\sigma}{\sqrt{n}}$

$\Rightarrow$

예제 264

$N(\mu,\ 3^2)$에서 얻은 크기 $n=16$인 확률표본의 평균 $\overline{X}=18.5$일 때 모평균 μ에 대한 95% 단측신뢰상한을 구하여라.

예제 265

실험용 생쥐의 특정자극에 대한 반응시간 X(초)는 $N(\mu,\ 4)$를 따른다고 알려져 있다. 랜덤추출한 25마리의 생쥐의 평균반응시간 $\overline{X}=6.2$(초)이었다. 평균반응시간 μ에 대한 95% 단측신뢰상한을 구하여라.

Thm (47): 모평균 μ의 구간추정②(σ^2을 모르는 경우)

모분산 σ^2이 알려져 있지 않은 정규모집단에서 얻은 크기 n인 확률표본의 평균과 표본분산을 $\overline{X}$와 s^2이라 하면

μ의 $100(1-\alpha)\%$ 신뢰구간은

$$\overline{X}-t_{\frac{\alpha}{2}(n-1)}\cdot\frac{s}{\sqrt{n}}\le\mu\le\overline{X}+t_{\frac{\alpha}{2}(n-1)}\cdot\frac{s}{\sqrt{n}} \quad \text{이다.}$$

여기서, $t_{\frac{\alpha}{2}(n-1)}$는 자유도 $n-1$인 t분포의 오른쪽 면적이 $\frac{\alpha}{2}$인 t값이다.

$\Rightarrow$

예제 266

전국학력평가에 응시한 수험생 중에서 수험생 9명을 임의로 선택하여 수리영역점수를 조사하였다. 그 결과 수험생은 평균 42점, 표준편차 7.5점이었다. 응시한 수험생 전체의 수리영역 점수가 정규분포를 이룬다면 95% 신뢰구간을 추정하여라.

※t분포표 $P(t\ge t_\alpha)=\alpha$

자유도 \ α	0.05	0.025
7	1.895	2.365
8	1.860	2.306
9	1.833	2.262
10	1.812	2.228

예제 267

시중에 유통되는 야채통조림에 포함되어 있는 무기질의 양 X는 정규분포를 따른다고 한다. 임의로 선택한 17개의 야채통조림에 함유되어 있는 무기질 양(mg)이 다음과 같을 때 이 야채통조림의 평균 무기질 함유량 μ에 대한 95%신뢰구간을 구하여라.

16 22 21 20 23 21 19 15 13 13 17 20 29 18 22 16 25

예제 268

한 마리의 젖소가 일정기간동안 생산해낸 우유의 우유지방의 양 X는 정규분포를 따른다고 한다. 한 농부가 20마리의 젖소를 임의로 선택조사하였더니 아래와 같은 우유지방측정치를 얻었다. μ에 대한 90%신뢰구간을 구하여라.

475 555 599 392 570 500 505 637 453 510
513 583 350 643 481 537 618 327 499 421

Thm (48) : 대표본 신뢰구간(σ^2을 모를 때)

표본 크기 n이 충분히 클 때($n \geq 30$일 때) 표본표준편차 s가 모표준편차 σ에 근사하므로 구간 $\left[\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}},\ \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]$을 근사신뢰구간으로 잡을 수 있다.

예제 269

어느 공장의 제품을 임의로 100개 추출하여 무게를 측정하였더니 평균 25, 표준편차 4.99이었을 때 모평균 μ에 대한 95% 신뢰구간을 구하여라.

예제 270

어느 회사의 컴퓨터 중 45대를 임의추출하여 디스크 드라이브 오류시간을 조사하였더니 평균 1762시간, 표준편차가 215시간이었다. 모평균 μ에 대한 90%신뢰구간을 구하여라.

Thm (49) : 모분산 σ^2에 관한 구간추정

정규분포에서 얻은 크기 n인 확률표본의 분산을 s^2이라 하면
σ^2의 $100(1-\alpha)\%$ 신뢰구간은

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}$$ 이다.

여기서, $\chi^2_{\frac{\alpha}{2}}$와 $\chi^2_{1-\frac{\alpha}{2}}$는 자유도 $n-1$인 카이제곱분포의 오른쪽 면적이 각각 $\frac{\alpha}{2}$, $1-\frac{\alpha}{2}$가 되는 값이다.

$\Rightarrow$

예제 271
어느 공장에서 생산되는 제품의 인장 강도의 변동을 조사하기 위해 임의로 제품 10개를 추출하여 조사한 결과 분산은 3.4이었다. 모분산 σ^2에 대한 95% 신뢰구간을 구하여라. (단, 인장강도의 변동은 정규분포를 따른다.)

예제 272
어떤 회사에서 판매하는 꽃씨의 무게를 조사하기 위해 10봉지를 임의로 추출하여 조사한 결과 표본분산은 0.286이었다. 꽃씨봉지의 무게의 분산 σ^2에 대한 95% 신뢰구간을 구하여라. (단, 무게는 정규분포를 따른다.)

예제 273
어떤 화초의 씨가 여무는데 걸리는 시간을 조사하기 위해 13개의 확률표본을 조사하였더니 표본분산 $s^2 = 128.41$이었다. σ^2에 대한 90% 신뢰구간을 구하여라.
(단, 시간은 정규분포를 따른다.)

예제 274
어느 회사에서 생산되는 통조림 캔을 임의로 10개 선택하여 내용물의 무게를 조사하였더니 분산은 4.3^2이었다. 모분산 σ^2에 대한 90% 신뢰구간을 구하여라.
(단, 무게는 정규분포를 따른다.)

예제 275

어떤 회사에서 판매되는 제품의 무게는 정규분포를 따른다고 한다. 10개의 제품을 임의로 추출하여 그 무게를 측정한 값이 다음과 같을 때 모든 제품의 무게의 분산에 대한 95% 신뢰구간을 구하여라.

46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2, 46.0

예제 276

어느 학교의 학생들의 IQ는 정규분포를 따른다고 한다. 36명의 학생들을 추출하여 IQ테스트를 하였을 때 불편분산값이 96이 나왔다. 이 때 σ^2에 대한 95% 신뢰구간을 구하여라.

Thm (50) : 모비율 p에 대한 구간추정

모비율이 p인 이항분포에서 충분히 큰 크기 n의 확률표본을 얻었을 때

표본비율 $\hat{p}=\dfrac{X}{n}$에 대하여 모비율 p의 $100(1-\alpha)\%$ 근사신뢰구간은

$$\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$ 이다.

단) $z_{\frac{\alpha}{2}}$는 오른쪽 면적이 $\dfrac{\alpha}{2}$인 z의 값이다.

$\Rightarrow$

예제 277

새로 개발한 신약의 치료율을 조사하기 위해 환자 300명에게 이 신약을 처방한 결과 250명이 효과가 있었다고 한다. 모비율 p에 대한 95% 신뢰구간을 구하여라.

예제 278

특정 부품의 10년 내구성에 대한 조사에서 295개를 조사하였더니 118개에서 결함이 발생하였다. 결함이 발생한 실제 부품의 비율에 대한 95% 신뢰구간을 구하여라.

예제 279

국회의원 선거에 출마한 한 후보자가 전체유권자에 대한 무작위 여론조사를 하였다. 조사결과 351명의 유권자 중 185명이 그의 지지자였다. 전체유권자 중 그의 지지자 비율에 대한 95% 신뢰구간을 구하여라.

Thm (51): 두 모평균 차의 추정(σ_1^2, σ_2^2을 아는 경우)

분산이 각각 σ_1^2, σ_2^2인 두 정규모집단에서 얻은 확률표본의 크기가 n_1과 n_2이고, 서로 독립인 이들 확률표본의 평균을 각각 $\overline{X}$, $\overline{Y}$라 하면 $\mu_1-\mu_2$의 $100(1-\alpha)\%$ 신뢰구간은

$$(\overline{X}-\overline{Y})-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1-\mu_2 \le (\overline{X}-\overline{Y})+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$ 이다.

여기서, $z_{\frac{\alpha}{2}}$는 오른쪽 면적이 $\frac{\alpha}{2}$인 z의 값이다.

$\Rightarrow$ 정규모집단에서 표본을 추출하는 위의 경우 이 신뢰구간은 정확한 구간이 된다. 그러나, 비록 비정규모집단이더라도 표본의 크기가 충분히 크면 중심극한정리에 의해 근사신뢰구간이 된다.

예제 280

동일조건으로 두 기계 A와 B의 시험가동을 각각 50번, 75번 한 결과 평균비용이 각각 36, 42으로 나타났다. (단위: 만원) 두 기계의 비용의 모평균 μ_A, μ_B의 차에 대한 95% 신뢰구간을 구하여라. (단, 각각 모표준편차는 6과 8이라고 알려져 있다.)

예제 281

A, B 두 제품의 평균수명의 차를 조사하려고 각각 15개씩 추출하여 조사한 결과 표본평균이 각각 480시간, 452시간이었다. 두 제품의 수명시간은 모표준편차가 각 $\sigma_1 = 34$, $\sigma_2 = 25$이고 정규분포를 따른다고 한다. 두 제품의 평균수명시간의 차 $\mu_1 - \mu_2$에 대한 95% 신뢰구간을 구하여라.

예제 282

두 전구의 수명시간 X와 Y는 각각 $\sigma_1^2 = 784$, $\sigma_2^2 = 627$인 정규분포를 따르고 X와 Y가 독립일 때 각각 56개, 57개의 표본에서 $\overline{X} = 937.4$, $\overline{Y} = 988.9$를 얻었다. $\mu_y - \mu_x$에 대한 90% 신뢰구간을 구하여라.

Def (48) : 등분산 σ^2의 합동추정량($\sigma_1^2 = \sigma_2^2 = \sigma^2$)

$$S_p^2 = \frac{\sum_{i=1}^{n_1}\left(X_i - \overline{X}\right)^2 + \sum_{i=1}^{n_2}\left(Y_i - \overline{Y}\right)^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

여기서, $S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_i - \overline{X}\right)^2$, $S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(Y_i - \overline{Y}\right)^2$ 이다.

※ $\chi^2 = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}$ 은 자유도 $n_1 + n_2 - 2$ 인 카이제곱분포를 따른다.

※ $T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ 은 자유도 $n_1 + n_2 - 2$ 인 t분포를 따른다.

Thm (52) : $\mu_1-\mu_2$의 신뢰구간($\sigma_1^2=\sigma_2^2$이며 모르는 경우)

모분산은 모르지만 같다고 볼 수 있는 두 정규모집단으로부터 서로 독립인 크기 n_1과 크기 n_2인 확률표본의 평균을 각각 $\overline{X}$, $\overline{Y}$라 하면 $\mu_1-\mu_2$의 $100(1-\alpha)\%$ 신뢰구간은

$$(\overline{X}-\overline{Y})-t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\le\mu_1-\mu_2\le(\overline{X}-\overline{Y})+t_{\frac{\alpha}{2}}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$ 이다.

여기서, S_p는 모표준편차 σ의 합동추정량이고 $t_{\frac{\alpha}{2}}$는 자유도 n_1+n_2-2인 t분포의 오른쪽 면적이 $\frac{\alpha}{2}$인 t값이다.

$\Rightarrow$

예제 283

두 모집단은 동일한 분산을 가지는 정규분포를 따른다고 한다. 각각 모집단에서 표본의 크기 12, 10인 확률표본을 얻었을 때 각각의 평균이 85, 81이고 각각의 분산이 16, 25일 때 두 모평균의 차 $\mu_1-\mu_2$의 90% 신뢰구간을 구하여라.

예제 284

A와 B 두 학과 학생들의 통계학 시험점수는 각각 같은 분산을 가지는 정규분포를 따른다고 한다. 이 때 분산은 모르는 값이다.
A학과 학생 9명의 확률표본에서 평균은 81.31, 분산은 60.76이 나왔고
B학과 학생 15명의 확률표본에서 평균은 78.61, 분산은 48.24이 나왔다면
$\mu_A - \mu_B$의 95% 신뢰구간을 구하여라.

예제 285

한강의 수질 오염을 조사하기 위해 하류에서 12개월, 상류에서 10개월 동안 수집된 표본에서 얻은 다양성 지수의 평균값 $\overline{X}=3.11$, $\overline{Y}=2.04$, 표준편차 $s_1=0.771$, $s_2=0.448$이었을 때 두 지역의 수질 오염의 모평균 차이 $\mu_1-\mu_2$의 90% 신뢰구간을 구하여라. (단, 두 지역의 수질오염은 모분산이 같은 정규분포를 따른다고 한다.)

〈참고〉 $\sigma_1^2 \neq \sigma_2^2$ 이며 모르는 경우

(방법①) : 소표본인 경우

$\mu_1 - \mu_2$ 의 $100(1-\alpha)\%$ 근사신뢰구간은

$$(\overline{X} - \overline{Y}) - t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{X} - \overline{Y}) + t_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$ 이다.

여기서, $t_{\frac{\alpha}{2}}$ 는 자유도가 다음과 같은 t 분포의 오른쪽 면적이 $\frac{\alpha}{2}$ 인 t 값이다.

자유도 $v = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$

(방법②) : 대표본인 경우

만약 두 분포의 크기가 클 때에는 비록 $\sigma_1^2 \neq \sigma_2^2$ 이며 모르는 경우에도

$\mu_1 - \mu_2$ 의 $100(1-\alpha)\%$ 근사신뢰구간은

$$(\overline{X} - \overline{Y}) - z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{X} - \overline{Y}) + z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$ 이다.

($\because$ Thm (51) 에서 n_1 과 n_2 가 충분히 크면 $s_1^2 \to \sigma_1^2$, $s_2^2 \to \sigma_2^2$ 이기 때문에)

예제 286

토플시험을 친 학생 중에서 임의로 남학생 75 명, 여학생 50 명을 추출하여 점수를 조사한 결과 남학생은 평균 82 점, 표준편차 8 점이었고 여학생은 평균 76 점, 표준편차 6 점이었다. $\mu_1 - \mu_2$ 의 95% 신뢰구간을 구하여라.

Thm (53) : 두 모비율의 차 $p_1 - p_2$의 신뢰구간

이항분포를 따르는 두 모집단에서 얻은 크기 n_1과 n_2인 확률표본의 성공횟수를 각각 X, Y 성공비율을 각각 $\hat{p_1}$, $\hat{p_2}$이라 하면 두 모비율의 차 $p_1 - p_2$의 $100(1-\alpha)\%$ 근사신뢰구간은

$$(\hat{p_1} - \hat{p_2}) - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} \leq p_1 - p_2 \leq (\hat{p_1} - \hat{p_2}) + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

이다. 여기서, $z_{\frac{\alpha}{2}}$는 표준정규분포의 오른쪽 면적이 $\frac{\alpha}{2}$인 값이다.

$\Rightarrow$

예제 287

어느 도시에 시장 선거에 출마한 후보가 A구에서 600명을 임의로 추출하여 조사한 결과 지지자가 320명이었고 B구에서 500명을 임의로 추출하여 조사한 결과 지지자가 220명이었다. 두 구의 지지율의 차 $p_1 - p_2$에 대한 95% 신뢰구간을 구하여라.

예제 288

어떤 공장에서 두 개의 생산라인에서 생산되는 제품의 불량률의 차이를 조사하기 위해 각 라인에서 생산되는 제품을 임의로 1500개, 2000개를 추출하여 조사한 결과 각각 불량품이 75개, 80개로 나타났다. 두 생산라인의 불량률의 차 $p_1 - p_2$에 대한 90% 신뢰구간을 구하여라.

Thm (54) : 두 모분산비(σ_1^2/σ_2^2)의 신뢰구간

두 정규모집단에서 독립적으로 추출한 크기 n_1과 n_2인 확률표본의 분산을 각각 s_1^2, s_2^2이라 하면 $\dfrac{\sigma_1^2}{\sigma_2^2}$의 $100(1-\alpha)\%$ 신뢰구간은

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\frac{\alpha}{2}}(n_1-1,\ n_2-1)} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2}{s_2^2}\cdot F_{\frac{\alpha}{2}}(n_2-1,\ n_1-1) \quad \text{이다.}$$

여기서 $F_{\frac{\alpha}{2}}(n_1-1,\ n_2-1)$는 자유도 $(n_1-1,\ n_2-1)$인 F분포의 오른쪽 면적이 $\dfrac{\alpha}{2}$인 값이고 $F_{\frac{\alpha}{2}}(n_2-1,\ n_1-1)$은 자유도 $(n_2-1,\ n_1-1)$인 F분포의 오른쪽 면적이 $\dfrac{\alpha}{2}$인 값이다.

$\Rightarrow$

예제 289

어느 백화점의 판매고는 정규분포를 따른다고 한다. 고졸 출신 26명을 임의추출하여 조사한 결과 $s_1^2 = 6000$, 대졸출신 16명을 임의추출하여 조사한 결과 $s_2^2 = 5000$이었을 때 모분산 비에 대한 95% 신뢰구간을 구하여라.

예제 290

두 정규모집단에서 각각 10개씩 임의추출한 확률표본의 분산이 $s_1^2 = 0.0215$, $s_2^2 = 0.0276$일 때 두 모분산비 $\dfrac{\sigma_1^2}{\sigma_2^2}$에 대한 95% 신뢰구간을 구하여라.

Def (49) : 우도함수(likelihood function)

확률변수 $X_1,\ X_2,\ \cdots\ ,\ X_n$의 결합확률밀도함수 $f(x_1,\ x_2,\ \cdots\ ,\ x_n\,;\theta)$를 모수 θ에 대한 함수로 볼 때 이를 **우도함수**라 한다.
즉, $L(\theta\,;x_1,\ x_2,\ \cdots\ ,\ x_n)=f(x_1,\ x_2,\ \cdots\ ,\ x_n\,;\theta)$이다.

$\Rightarrow$ 확률변수 $X_1,\ X_2,\ \cdots\ ,\ X_n$이 서로 독립이고 확률밀도함수 $f(x\,;\theta)$에서 얻은 확률표본이라 하면 $L(\theta\,;x_1,\ x_2,\ \cdots\ ,\ x_n)=f(x_1\,;\theta)f(x_2\,;\theta)\ \cdots\ f(x_n\,;\theta)=\prod_{i=1}^{n}f(x_i\,;\theta)$이다.

Def (50) : 최대우도 추정량(=최우 추정량)

확률표본의 우도함수 $L(\theta\,;x_1,\ x_2,\ \cdots\ ,\ x_n)$를 최대로 하는 θ의 값을 $\hat{\theta}$이라 할 때 $\hat{\theta}$을 모수 θ의 **최대우도 추정량**이라 한다.

$\Rightarrow$

예제 291

확률질량함수가 $f(x\,;\mu)=\dfrac{e^{-\mu}\mu^x}{x!}$, $x=0,\ 1,\ 2,\ \cdots$ 인 포아송 분포에서 독립적으로 얻은 확률 표본 $X_1,\ X_2,\ \cdots\ ,\ X_n$에 대하여 μ의 최대우도 추정량을 구하여라.

예제 292

$X_1,\ X_2,\ \cdots\ ,\ X_n$을 정규분포 $N(\mu,\ \sigma^2)$으로부터 얻은 확률표본이라 할 때 μ와 σ^2의 최대우도추정량을 구하여라.

예제 293

$X_1,\ X_2,\ \cdots\ ,\ X_n$을 성공확률이 p인 베르누이 분포에서 얻은 확률표본이라 할 때 p의 최대추정량을 구하여라.

예제 294

X_1, X_2, $\cdots$, X_n이 확률밀도함수 $f(x\,;\theta)=\frac{1}{\theta}e^{-\frac{x}{\theta}}$, $0<x<\infty$를 가진 지수분포에서 추출된 확률표본이라 할 때 θ의 최우 추정량을 구하여라.

예제 295

신약의 효능을 실험하기 위해 10마리의 암에 걸린 실험용 쥐에 투여 후 생존시간을 기록한 결과가 다음과 같다. 생존시간 X가 지수분포를 따른다면 평균생존시간 θ에 대한 최우추정량을 구하여라.

14, 17, 27, 18, 12, 8, 22, 13, 19, 12

제 9 장　검 정

Def (51) : 통계적 가설 검정

표본에서 얻은 사실을 근거로 하여 모집단에 대한 가설이 맞는지 통계적으로 검정하는 분석방법

$$\text{가설}\begin{cases} \bullet\ \text{귀무가설}(H_0) \sim \text{직접 검정 대상이 되는 가설} \\ \bullet\ \text{대립가설}(H_1 \text{ or } H_a) \sim \text{귀무가설이 기각될 때 받아들여지는 가설} \end{cases}$$

예제 296

어떤 회사의 계약직의 작년 평균월급은 170만원, 표준편차는 30만원 이었다고 한다. 올해는 그보다 높을 것이라고 생각하여 임의로 100명을 골라 평균 월급을 조사하였더니 175만원이었다. 월급이 170만원 이상이라고 할 수 있는지를 유의수준 $\alpha = 0.05$에서 검정하라.

Def (52) : 유의수준과 임계값(기각 역의 경계값)

임계값이란 주어진 유의수준에서 귀무가설의 채택과 기각에 관련된 의사를 결정할 때 그 기준이 되는 점이다. 여기서 유의수준은 귀무가설이 옳음에도 기각할 오류 (α − 오류)이다.

$\Rightarrow$

Def (53): 가설검정에서의 오류

(1) **제 1종 오류(α−오류)**

귀무가설이 참임에도 이를 기각하는 오류 ⇨ 유의수준

(2) **제 2종 오류(β−오류)**

귀무가설이 거짓임에도 이를 기각하지 않는 오류(즉, 채택하는 오류)

※$1-\beta$를 **통계적 검정력**이라 한다. ⇨ 일반화: 검정력 함수 $\pi(\theta)$로 둔다.

$\Rightarrow$

※ 두 오류의 확률 α와 β의 관계

① 주어진 표본크기 n 아래서는 α와 β 동시에 줄일 수 없다.

즉, α를 작게 하면 β가 커지고 β를 작게 하면 α가 커진다.

② α와 β를 동시에 줄이려면 표본크기 n을 증가시켜야 한다.

예제 297

X_1, X_2, $\cdots$, X_{25}를 정규분포 $N(\mu,\ 10^2)$에서 얻은 확률표본이라 하자.
귀무가설 H_0 : $\mu=100$, 대립가설 H_1 : $\mu=105$를 검정할 때 기각역 $\{\overline{X_{25}}\geq 104\}$일 때 제 1종 오류를 범할 확률 α와 제 2종 오류를 범할 오류 β를 각각 구하여라.

예제 298

예제 297에서 $\mu=105$일 때 검정력을 구하고 $\mu>100$인 모든 점에서의 검정력 즉, 검정력 함수 $\pi(\mu)$를 구하여라.

Thm (55) : 모평균에 대한 가설검정(σ^2을 아는 경우)

모분산 σ^2을 알고 있을 때 $N(\mu,\ \sigma^2)$의 모평균 μ에 대한 가설검정에서

① 검정통계량 $Z=\dfrac{\overline{X}-\mu_0}{\dfrac{\sigma}{\sqrt{n}}}$로 둔다.

② 유의수준 α에 따른 기각역

$|z| \geq z_{\frac{\alpha}{2}}$, $z \geq z_\alpha$ 또는 $z \leq -z_\alpha$라 둔다.

예제 299

내용물의 무게가 350g으로 표시된 참치 캔 생산공정에서 평균무게가 360g이 되도록 공정을 관리하고 있다. 이를 확인하기 위해 30개의 표본을 추출하여 조사하였더니 평균 356g이 나왔다. 내용물의 무게가 표준편차 10g인 정규분포를 따른다고 할 때 생산공정에 이상이 있다고 할 수 있는지를 유의수준 $\alpha=0.05$로 검정하라.

※ 유의수준에 따른 임계값

α	양측 검정		단측 검정	
	$-z_{\frac{\alpha}{2}}$	$z_{\frac{\alpha}{2}}$	$-z_\alpha$	z_α
0.01	-2.57	2.57	-2.33	2.33
0.02	-2.33	2.33	-2.05	2.05
0.05	-1.96	1.96	-1.645	1.645
0.10	-1.645	1.645	-1.28	1.28

예제 300

우리나라 사람의 평균수명을 알아보기 위해 사망자 100명을 표본으로 추출하여 조사하였더니 평균 71.8년이었다. $\sigma = 8.9$년이라 가정할 때 현재의 수명은 70년보다 길다고 할 수 있는지를 유의수준 $\alpha = 0.05$에서 검정하라.

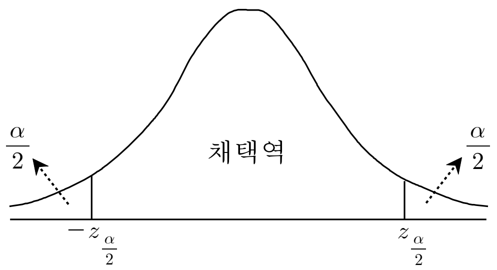
예제 301

어느 백화점에서 판매원에 대한 재교육을 실시한 후에 판매고에 대한 조사를 하고자 임의로 9일을 선택하여 조사하였더니 평균 판매고는 1100(단위:천원) 이었다. 평균판매고가 1000을 초과하면 재교육이 효과가 있는 것으로 간주할 때 재교육이 효과가 있었는지 유의수준 $\alpha = 0.01$로 검정하여라. (단, 판매고는 $\sigma = 100$인 정규분포를 따른다.)

예제 302

낚시줄을 새로 개발한 업자가 신제품의 평균인장강도 8kg, 표준편차 0.5kg이라 주장하고 있다. 낚시줄 50개를 표본으로 추출하여 조사한 결과 평균인장강도는 7.8kg으로 나타났다. 유의수준 $\alpha = 0.01$로 검정하여라.

※ 검정과 구간추정과의 관계 (μ의 예)



$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

μ의 $100(1-\alpha)\%$ 신뢰구간에 μ_0가 포함된다는 것은

"μ_0가 채택역 안에 있으므로 H_0가 기각되지 않는다."

와 같은 말이다.

$$\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

따라서, "H_0를 기각한다."와 "μ의 $100(1-\alpha)\%$ 신뢰구간에 μ_0가 포함되지 않는다."와 같은 말이다.

Thm (56) : 모평균에 대한 가설검정(σ^2 을 모르는 경우)

(1) 대표본($n \geq 30$) 일 때

① 검정통계량 $Z = \dfrac{\overline{X} - \mu_0}{\dfrac{s}{\sqrt{n}}}$

② 유의 수준 α 에 따른 기각역

$|z| \geq z_{\frac{\alpha}{2}}$, $z \geq z_{\alpha}$ 또는 $z \leq -z_{\alpha}$ 이다.

(2) 소표본($n < 30$) 일 때

① 검정통계량 $T = \dfrac{\overline{X} - \mu_0}{\dfrac{s}{\sqrt{n}}}$ ($\because$ 자유도 $n-1$인 t분포)

② 유의수준 α 에 따른 기각역

$|T| \geq t_{\frac{\alpha}{2}}(n-1)$ 또는 $T \geq t_{\alpha}(n-1)$ 또는 $T \leq -t_{\alpha}(n-1)$

예제 303

지난 한 해 동안 사망한 사람 중 100 명을 추출하여 조사하였더니 평균수명이 71.8 세이고 표준편차는 8.9 세였다. 이 자료에 의하여 평균수명이 70 세 이상이라고 할 수 있는지 유의수준 $\alpha = 0.05$ 에서 검정하라.

예제 304

어느 간장제조회사에서 생산되는 간장의 평균용량은 200 ml 로 표시되어 있다. 임의로 20 개를 추출하여 조사한 결과 평균은 200.2 ml, 표준편차는 0.32 ml 이었다. 이 회사에서 생산되는 간장의 평균용량이 200 ml 라 할 수 있는지 유의수준 $\alpha = 0.05$ 에서 검정하여라.

예제 305

A회사는 그 공장에서 인근 강으로 내보내는 폐기물의 산소소비량이 예전의 평균 $\mu = 500$ (p.p.m) 보다 낮아졌다고 주장한다. $n = 25$ 일간 조산한 결과 $\overline{X} = 308.8$, $s = 115.15$ 로 나왔다. $H_0 : \mu = 500$, $H_1 : \mu < 500$ 라 두고 이 주장을 유의수준 $\alpha = 0.01$ 에서 검정하여라.

Thm (57): 모비율에 대한 가설검정

① 검정통계량 $Z=\dfrac{\hat{p}-p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

② 유의수준 α에 따른 기각역

$|z| \geq z_{\frac{\alpha}{2}}$, $z \geq z_{\alpha}$ 또는 $z \leq -z_{\alpha}$

예제 306

동전이 정상적인가를 검정하기 위해 100번을 던져본 결과 앞면 40번, 뒷면 60번이 나왔다. 이 동전은 정상인가를 유의수준 $\alpha=0.05$로 검정하여라.

예제 307

모 불량률 $p_0=0.082$인 어느 공정에서 임의로 160개의 제품을 추출하여 검사한 결과 불량품이 15개 나왔다. 불량률이 달라졌다고 할 수 있는지를 유의수준 $\alpha=0.05$로 검정하여라.

예제 308

현재 사용되는 신경안정제의 효과는 60%라고 한다. 신경과민환자 100명을 임의 추출하여 새로 개발한 약을 처방한 결과 70명이 효과를 보았다고 한다. 새로 개발한 약이 기존 약보다 효과가 더 크다고 할 수 있는지 유의수준 $\alpha = 0.05$로 검정하여라.

예제 309

어느 공장에서 생산되는 제품 100개를 임의 추출하여 조사한 결과 불량품이 3개 나왔다. 이 공장에서 생산되는 제품의 불량률은 2%라고 할 수 있는지를 유의수준 $\alpha = 0.05$로 검정하여라.

Thm (58): 모분산에 대한 가설검정

① 검정통계량 $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$ (자유도 $(n-1)$인 카이제곱분포를 따른다.)

② 유의수준 α에 따른 기각역

$\chi^2 \geq \chi^2_{\frac{\alpha}{2}}(n-1)$ 또는 $\chi^2 \leq \chi^2_{1-\frac{\alpha}{2}}(n-1)$ (양측검정)

$\chi^2 \geq \chi^2_{\alpha}(n-1)$ (우측단측검정), $\chi^2 \leq \chi^2_{1-\alpha}(n-1)$ (좌측단측검정)

예제 310

정규분포를 이루는 어떤 모집단의 분산이 20이라 가정하고 30개의 표본을 추출하여 분산 s^2이 32이었다. 유의수준 5%로써 양측검정으로 가설을 검정하라.

예제 311

통계학자 P씨는 어느 대학의 학생들의 IQ지수의 분산이 $\sigma^2 = 100$이라고 주장했다. 임의로 23명의 학생들을 추출하여 조사한 결과 불편추정치 $s^2 = 172$이 나왔다. P씨의 주장을 유의수준 $\alpha = 0.05$로 검정하여라.

예제 312

어느 전구회사에서 작업공정을 일부 변경하였다. 새로운 공정에서 생산된 전구 30개를 추출하여 조사한 결과 분산이 105시간이었다. 이전 전구의 수명은 분산이 120시간인 정규분포를 따랐다면 새로운 공정에서 전구수명의 분산은 이전보다 작아지는지를 유의수준 5%로 검정하여라.

예제 313

정부는 우리나라 중소기업의 종업원의 임금의 표준편차는 $\sigma = 500$원인 정규분포로 비교적 안정적이라고 발표했다. 그러나 많은 사람들은 실제 표준편차는 이보다 커서 임금격차가 심하다고 생각하며, 또 어떤 사람들은 전국의 중소기업의 임금수준은 비슷할 것으로 생각하여 표준편차가 더 작을 것이라고 생각하고 있다. 이를 검정하려고 51명을 임의 추출하여 조사한 결과 표준편차 $s = 700$원이었다. 정부발표를 유의수준 $\alpha = 0.1$로 검정하여라.

Thm (59) : 두 모평균의 차에 대한 가설검정(σ_1^2, σ_2^2을 알고있는 경우)

① 검정통계량 $Z=\dfrac{(\overline{X}-\overline{Y})-\delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$ $(\because\ \delta_0=\mu_1-\mu_2)$

② 유의수준 α에 따른 기각역

$|z|\geq z_{\frac{\alpha}{2}}$ 또는 $z\geq z_\alpha$ 또는 $z\leq -z_\alpha$

예제 314

표준편차가 $\sigma_1=4.8$, $\sigma_2=3.1$인 두 정규모집단에서 각각 임의로 표본크기 $n_1=16$, $n_2=25$을 추출하여 구한 평균 $\overline{X_1}=78$, $\overline{X_2}=70$이었다. 두 모평균이 같은지 유의수준 5%에서 검정하여라.

예제 315

어느 대학의 통계학과와 수학과 학생들의 통계학 중간시험 성적은 두 과 모두 표준편차 $\sigma=5$점인 정규분포를 이룬다고 한다. 이 두 과 학생들을 각각 $n_1=50$명, $n_2=50$명씩 임의추출하여 조사한 결과 $\overline{X_1}=83$점, $\overline{X_2}=86$점이었다.

이 두 학과 학생들의 통계학 성적의 평균의 차이가 있는지를 유의수준 5%에서 검정하여라.

예제 316

어느 회사는 신입사원을 대상으로 사전교육을 시키는 것이 효과적이라고 판단해서 10명을 임의 추출하여 사전교육을 시킨 결과 특정업무처리시간이 평균 12.1시간 걸렸고, 16명을 임의추출하여 사전교육없이 특정업무를 처리케한 결과 평균 14.2시간 걸렸다. 이 결과로부터 사전교육은 효과가 있다고 할 수 있는지를 유의수준 $\alpha = 0.05$에서 검정하여라. (단, 각 모집단은 $\sigma_1 = \sigma_2 = 4$인 정규분포이다.)

예제 317

어느 공장에서 남자직원 30명과 여자 직원 55명이 개인당 제품 하나를 만드는데 걸리는 시간을 조사한 결과 각각 평균이 7분, 5분이었다. 모집단의 $\sigma_1 = \sigma_2 = 2$일 때 과연 여자직원이 남자직원보다 더 빨리 제품을 만든다고 할 수 있는지를 5% 유의수준에서 검정하여라.

Thm (60) : 두 모평균의 차에 대한 가설검정(σ_1^2, σ_2^2을 모르고 대표본인 경우)

① 검정통계량 $Z=\dfrac{(\overline{X}-\overline{Y})-\delta_0}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}$ $(\because\ \delta_0=\mu_1-\mu_2)$

② 유의수준 α에 따른 기각역

$|z|\geq z_{\frac{\alpha}{2}}$ 또는 $z\geq z_{\alpha}$ 또는 $z\leq -z_{\alpha}$

예제 318

어느 고등학교에서 새로운 교육방식을 도입하고 이것이 성적향상에 도움이 되는지 조사하기 위해 $n_1=62$명을 임의추출하여 새로운 교육방식으로 교육 후 시험을 친 결과 평균 74점, 표준편차 16점이 나왔고, $n_2=60$명을 임의 추출하여 기존방식으로 교육후 시험을 친 결과 평균 66점, 표준편차 12점이 나왔다. 새로운 교육방식이 효과가 있는지를 유의수준 $\alpha=0.05$에서 검정하여라.

Thm (61): 두 모평균의 차에 대한 가설검정($\sigma_1^2 = \sigma_2^2$이며 모르고 소표본인 경우)

① 검정통계량 $T = \dfrac{\left(\overline{X_1} - \overline{X_2}\right) - \delta_0}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

여기서 $s_p = \sqrt{\dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$

② 유의수준 α에 따른 기각역

$|T| \geq t_{\frac{\alpha}{2}}(n_1 + n_2 - 2)$ 또는 $T \geq t_{\alpha}(n_1 + n_2 - 2)$ 또는 $T \leq -t_{\alpha}(n_1 + n_2 - 2)$

예제 319

독립인 두 정규모집단에서 다음과 같은 자료가 추출되었다.

$n_1 = 11$, $\overline{X_1} = 75$, $s_1 = 6.1$

$n_2 = 14$, $\overline{X_2} = 60$, $s_2 = 5.3$

모분산이 동일하다면 두 모평균이 같은지를 5%로 유의수준에서 검정하여라.

예제 320

어느 외국어고등학교에서 두 종류의 방법으로 학생들이 영어 단어를 암기하게 가르친 다음 각각의 방법에 대한 단어암기개수 결과에 대한 다음과 같은 자료를 얻었다.

방법 1 : 5, 6, 8, 7, 3, 4, 4, 2, 4, 7 $(n_1 = 10)$

방법 2 : 6, 4, 4, 8, 6, 5, 9, 6, 5, 7 $(n_2 = 10)$

위의 자료를 이용하여 두 가지 방법의 효과가 차이가 있는지를 유의수준 $\alpha = 0.1$ 에서 검정하라. (단, 두 모집단의 분산은 같고, 정규분포를 이룬다.)

예제 321

자동차를 생산하는 두 회사의 자동차를 각각 6대, 10대씩 표본추출하여 성능에 대한 종합점수에 대한 결과는 다음과 같다.

A사 자동차 : 77, 79, 84, 83, 82, 90

B사 자동차 : 94, 85, 90, 95, 90, 92, 75, 80, 79, 85

판매가격을 확인해보니 B사 자동차의 가격이 A사보다 비싼 것으로 나타났다. 과연 B사 자동차가 A사 자동차보다 우수한지를 유의수준 $\alpha = 0.05$ 에서 검정하여라.
(단, 두 모집단의 분산은 같고 정규분포를 이룬다.)

✔ 강의에서는 Thm (61)로 되어 있습니다.

Thm (62): 대응표본인 경우의 모평균차의 가설검정

각 조가 동질적인 2개의 실험단위의 쌍으로 이루어진 대응표본인 경우에 두 모평균의 차 $\mu_1 - \mu_2 = \delta_0$에 대한 가설검정에서

① 검정통계량 $T = \dfrac{\overline{D} - \delta_0}{\dfrac{s_d}{\sqrt{n}}}$ (자유도 $n-1$인 t분포)

여기서, $\overline{D} = \dfrac{\sum(X_1 - X_2)}{n} = \dfrac{\sum D_i}{n}$, $s_d^2 = \dfrac{\sum(D_i - \overline{D})^2}{n-1}$ 이다.

② 유의수준 α에 따른 기각역

$|T| \geq t_{\frac{\alpha}{2}}(n-1)$ 또는 $T \geq t_{\alpha}(n-1)$ 또는 $T \leq -t_{\alpha}(n-1)$

※ 대응표본의 예

A와 B 두 회사의 진통제효과를 알아보기 위해 이 진통제를 투여 후 평균진통시간의 차를 검정하는 경우 만약 독립적 표본추출을 한다면 환자간의 차이로 인한 진통효과의 차이가 영향을 미쳐 정확한 검정이 될 수 없다. 이러한 환자의 차이를 배제하기 위하여 동일한 환자에게 두 회사의 진통제를 각각 투여한 후 진통시간의 차를 비교한 테이터를 사용하면 된다.

예제 322

어느 기업에서 사원에 대한 직업교육을 실시한 후 능률향상에 효과가 있는지를 조사하려고 한다. 사원 10명을 임의추출하여 조사한 결과 직업교육 실시 전과 후의 업무능력의 차를 조사한 데이터가 다음과 같을 때 직업교육을 실시하기 전과 후에 업무능력의 차가 있는지를 5% 유의수준에서 검정하여라.

	1	2	3	4	5	6	7	8	9	10
실시전	23	45	35	28	27	29	25	30	41	33
실시후	27	43	31	42	34	38	30	37	36	44

예제 323

특정약은 사용자의 혈압을 저하시키는가를 조사하려고 15명의 환자를 대상으로 평상시 혈압을 측정한 뒤 이들에게 이 약을 투여 후 혈압을 측정한 결과가 다음과 같다. 이 약을 사용 후 평균혈압이 저하되었다고 할 수 있는지를 유의수준 $\alpha=0.05$에서 검정하여라. (단, x_1은 사용 전, x_2는 사용 후)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	70	80	72	76	76	76	72	78	82	64	74	92	74	72	84
x_2	68	72	62	70	58	56	68	52	64	72	64	60	70	72	74
$d=x_2-x_1$	-2	-8	-10	-6	-18	-20	-4	-26	-18	8	-10	-32	-4	0	-10

✔ 강의에서는 Thm (62)로 되어 있습니다.

Thm (63) : 두 모비율 차의 가설검정($p_1 = p_2$ 의 가설검정)

① 검정통계량 $Z = \dfrac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

$\left(\because\ \hat{p}\text{는 모수 } p\text{의 합동추정값 } \hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2}\text{이다.}\right)$

② 유의수준 α에 따른 기각역

$|z| \geq z_{\frac{\alpha}{2}}$ 또는 $z \geq z_\alpha$ 또는 $z \leq -z_\alpha$

$\Rightarrow$

예제 324

어떤 두 도시 인접지역에 물류단지를 건설하려고 한다. 이에 대한 두 도시 주민들을 대상으로 찬반투표를 실시하여 지지율 사이에 차이가 있는지를 알아보려고 한다. 만일 A도시 주민 200명을 임의추출하여 조사한 결과 120명이 찬성하고 B도시 주민 500명을 임의추출하여 조사한 결과 240명이 찬성했다고 한다.

A시의 지지율이 B시보다 높다고 할 수 있는지 유의수준 $\alpha = 0.05$로 검정하라.

예제 325

도지사 선거에 출마한 후보자가 도시와 시골의 지지율 차를 알아보기 위해 도시에서 500명, 시골에서 300명을 임의추출하여 조사한 결과 도시에서 280명 찬성, 시골에서 180명 찬성을 했다. 도시와 시골간에 지지율의 차이가 있는지를 유의수준 $\alpha = 0.05$ 에서 검정하여라.

예제 326

어느 공장의 A와 B 두 팀에서 생산한 제품의 불량률의 차이를 알아보려고 임의로 표본을 추출하여 조사한 결과 A팀에서는 105개 중 16개가 불량이었고, B팀에서는 96개 중 20개가 불량인 것으로 나타났다. A팀과 B팀의 불량률에 차이가 있는지를 유의수준 5%에서 검정하여라.

Thm (64) : 두 모분산비에 대한 가설검정($\sigma_1^2=\sigma_2^2$의 검정)

① 검정통계량 $F=\dfrac{s_1^2}{s_2^2}$

$$\left(\because\ F=\frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{\sigma_2^2}} \text{ 은 자유도 } (n_1-1,\ n_2-1)\text{인 } F\text{분포를 따른다.} \right) \leftarrow \text{Thm (64)}$$

② 유의수준 α에 따른 기각역

$F \geq F_{\frac{\alpha}{2}}(n_1-1,\ n_2-1)$ 또는 $F \leq F_{1-\frac{\alpha}{2}}(n_1-1,\ n_2-1)$ $\Rightarrow$ 양측검정

$F \geq F_{\alpha}(n_1-1,\ n_2-1)$ $\Rightarrow$ 우측단측검정

$F \leq F_{1-\alpha}(n_1-1,\ n_2-1)$ $\Rightarrow$ 좌측단측검정

$\Rightarrow$

예제 327

두 개의 모집단 $N(\mu_1,\ \sigma_1^2)$, $N(\mu_2,\ \sigma_2^2)$에서 $n_1=10$, $n_2=9$개의 표본을 구하여 각각 $\overline{x_1}=17.2$, $s_1^2=1.8$, $\overline{x_2}=14.7$, $s_2^2=8.7$을 얻었다. 모분산이 같은지를 유의수준 5%에서 검정하여라.

예제 328

A와 B 두 기계에서 생산되는 $1l$병 음료의 내용물의 변동에 차이가 있는지를 조사하려고 각각의 기계에서 임의로 20병씩 추출하여 조사한 결과 A 기계에서는 $\overline{x_1}=1013.5$, $s_1^2=39$이고 B기계에서는 $\overline{x_2}=1009.7$, $s_2^2=26.12$이었다. 내용물의 변동에 차이가 있다고 말할 수 있는지를 유의수준 $\alpha=0.05$에서 검정하여라.

예제 329

두 강재의 마모량을 비교하기 위한 실험이 실시되었다. 강재 1과 강재 2에 대하여 각각 12번, 10번 실험한 결과 강재 1의 마모량은 평균 85, 표준편차 4이었고 강재 2의 마모량은 평균 81, 표준편차 5이었다. 두 모분산이 같다고 말할 수 있는지를 유의수준 $\alpha = 0.1$에서 검정하여라.

예제 330

어느 공장에서 생산되는 제품 하나를 만드는 데 걸리는 시간을 남녀간에 차이가 있는지를 알아보는 실험이 실시되었다. 경험적으로 걸리는 시간은 남녀 모두 정규분포를 따르고 여성의 분산이 남성의 분산보다 작다고 한다.

11명의 남자와 16명의 여자를 임의추출하여 조사한 결과 각각 $s_1 = 6.1$, $s_2 = 5.3$이 나왔다. $\sigma_1^2 = \sigma_2^2$인지를 유의수준 5%로 검정하여라.

〈심화 특강 ④〉

검정력 함수 $\pi(\theta)$ ⇨ H_1이 사실일 때 H_0를 기각할 확률 $(1-\beta)$

ex 1)

새로운 교수법으로 수리통계학을 배운 학생들의 점수분포가 정규분포 $N(\mu,\ 10^2)$을 따른다고 하고 $X_1,\ X_2,\ \cdots,\ X_{25}$은 이 모집단에서 얻은 확률표본이다.

기존교수법으로 배운 학생들의 평균점수가 60점일 때 $H_0 : \mu = 60$(무변화)

$H_1 : \mu > 60$에 대하여 검정하려고 한다. 기각역 $\left\{\overline{X_{25}} \geq 62\right\}$일 때 $\mu \geq 60$인 모든 값에 대하여 검정력 함수 $\pi(\mu)$를 구하여라.

ex 2) 표본크기 n를 증가시키므로 α와 β의 감소

ex 1)에서 $n=100$이고 유의수준 $\alpha=0.05$인 검정에서 기각역을 구하고 $\mu=65$일 때 β를 구하여라.

ex 3) 검정력 함수를 이용한 표본크기의 결정

ex 1)에서 $P(\overline{X} \geq c \mid \mu=60) = 0.025$, $\mu=65$일 때 $\beta=0.05$이려면 표본크기 n을 얼마로 잡아야 하는가?

최강력 검정법 (=최량기각역)

Def : 최강력 검정법(most powerful test)

귀무가설 $H_0 : \theta = \theta_0$, 대립가설 $H_1 : \theta = \theta_1$ 에 대한 기각역 C^* 가 다음 (1), (2)를 만족할 때 유의수준이 α 인 **최강력검정법의 기각역**이라 한다.
π^* 를 기각영역 C^* 에 대한 검정력 함수라 할 때

(1) $\pi^*(\theta_0) = \alpha$ 이다.

(2) 유의수준이 α 인 임의의 기각영역 C에 대한 검정력 함수를 π 라 하면 $\pi^*(\theta_1) \geq \pi(\theta_1)$ 이다.

ex 4)
귀무가설과 대립가설 하에서 각각 확률밀도함수 $f(x \mid \theta_0)$ 와 $f(x \mid \theta_1)$ 이 다음과 같을 때 유의수준 $\alpha = 0.05$ 인 최강력 검정법(=최량기각역)을 구하여라.

X	1	2	3	4	5	6
$f(x \mid \theta_0)$	0.01	0.02	0.02	0.05	0.10	0.80
$f(x \mid \theta_1)$	0.03	0.05	0.05	0.10	0.20	0.57

ex 5)

$H_0 : \theta = \frac{1}{2}$ $H_1 : \theta = \frac{3}{4}$ 에 대한 유의수준 $\alpha = \frac{1}{32}$ 인 최량기각역을 다음 표에서 구하여라.

x	0	1	2	3	4	5
$f\left(x \,\middle|\, \frac{1}{2}\right)$	$\frac{1}{32}$	$\frac{5}{32}$	$\frac{10}{32}$	$\frac{10}{32}$	$\frac{5}{32}$	$\frac{1}{32}$
$f\left(x \,\middle|\, \frac{3}{4}\right)$	$\frac{1}{1024}$	$\frac{15}{1024}$	$\frac{90}{1024}$	$\frac{270}{1024}$	$\frac{405}{1024}$	$\frac{243}{1024}$
$f\left(x \,\middle|\, \frac{1}{2}\right) \Big/ f\left(x \,\middle|\, \frac{3}{4}\right)$	32	$\frac{32}{3}$	$\frac{32}{9}$	$\frac{32}{27}$	$\frac{32}{81}$	$\frac{32}{243}$

또, 유의수준 $\alpha = \frac{6}{32}$ 인 최량기각역을 H_0 가 참일 때의
부분집합 $\{x \mid x = 0, 1\}$, $\{x \mid x = 0, 4\}$, $\{x \mid x = 1, 5\}$, $\{x \mid x = 4, 5\}$ 에서
최량기각역을 골라라.

〈심화 특강 ⑤〉

Thm : 네이만-피어슨 정리(크기 α인 최량기각역에 대한 충분조건)

확률표본 $X_1,\ X_2,\ \cdots,\ X_n$의 결합확률밀도함수가 $f(x_1,\ x_2,\ \cdots,\ x_n\,;\theta)$ 로 주어졌다고 하자. 다음의 세 가지 조건

① $P\left[(X_1,\ X_2,\ \cdots,\ X_n)\in C^* \mid \theta_0\right]=\alpha$

② $(x_1,\ x_2,\ \cdots,\ x_n)\in C^*$ 에 대하여 우도비 $\dfrac{L(\theta_0)}{L(\theta_1)}\le k$ (k는 양의 상수)

③ $(x_1,\ x_2,\ \cdots,\ x_n)\in C^{*\prime}$ 에 대하여 우도비 $\dfrac{L(\theta_0)}{L(\theta_1)}\ge k$ (k는 양의 상수)

를 만족할 때 C^*는 $H_0 : \theta=\theta_0$와 $H_1 : \theta=\theta_1$에 대해 크기 α인 최량기각역이다.

$\Rightarrow$ $X=(X_1,\ X_2,\ \cdots,\ X_n)$이라 두면

$P[X\in A \mid \theta]=\int_A \cdots \int f(x_1,\ x_2,\ \cdots,\ x_n\,;\theta)\,dx_1dx_2\cdots dx_n$ 이므로

$A\subset C^*$ 이면

$\int_A\cdots\int f(x_1,\ x_2,\ \cdots,\ x_n\,;\theta_0)\,dx_1\cdots dx_n \le$

$\int_A\cdots\int kf(x_1,\ x_2,\ \cdots,\ x_n\,;\theta_1)\,dx_1\cdots dx_n$

이므로 $P\left[X\in A \mid \theta_0\right]\le kP\left[X\in A \mid \theta_1\right]$ 이 성립한다.

같은 방법으로 $A\subset (C^*)^c$ 이면

$P\left[X\in A \mid \theta_0\right]\ge kP\left[X\in A \mid \theta_1\right]$ 이 성립한다.

임의의 기각영역 C에 대하여 $C^*=(C^*\cap C)\cup(C^*\cap C^c)$, $C=(C\cap C^*)\cup\left(C\cap(C^*)^c\right)$ 로 표현된다.

그런데 기각영역 C^*와 C에 대한 검정력 함수는 각각

$\pi^*(\theta)=P\left[X\in C^*\cap C \mid \theta\right]+P\left[X\in C^*\cap C^c \mid \theta\right]$

$\pi(\theta)=P\left[X\in C\cap C^* \mid \theta\right]+P\left[X\in C\cap(C^*)^c \mid \theta\right]$ 이므로

검정력 차이 $\pi^*(\theta_1)-\pi(\theta_1)=P\left[X\in C^*\cap C^c \mid \theta_1\right]-P\left[X\in C\cap(C^*)^c \mid \theta_1\right]$

$$\ge \frac{1}{k}\left\{P\left[X\in C^*\cap C^c \mid \theta_0\right]-P\left[X\in C\cap(C^*)^c \mid \theta_0\right]\right\}$$

$$=\frac{1}{k}\left\{\pi^*(\theta_0)-\pi(\theta_0)\right\}=0$$

$\therefore$ $\pi^*(\theta_1)\ge\pi(\theta_1)$ 이므로 C^*는 크기 α인 최량기각역이다.

ex 6)

X_1, X_2, $\cdots$, X_n이 정규분포 $N(\mu,\ 36)$으로부터 구한 확률 표본이라 한다.
이제 귀무가설 H_0 : $\mu = 50$과 대립가설 H_1 : $\mu = 55$를 검정하기 위한 최량기각역을 찾아보자.

ex 7)

X_1, X_2, $\cdots$, X_n을 평균 λ인 포아송 분포에서 얻은 크기 n인 확률표본이라 하자.

H_1 : $\lambda = 5$에 대한 H_0 : $\lambda = 2$를 검정하기 위한 최량기각역을 구하라.

〈심화 특강 ⑥〉

ex 8)

$X_1,\ X_2,\ \cdots,\ X_n$ 은 정규분포 $N(\mu,\ 1)$ 으로부터 얻은 확률표본이다.

$H_0 : \mu = \mu_0 \quad H_1 : \mu = \mu_1\,(> \mu_0)$ 을 검정하기 위한 최량기각역을 구하여라.

Def : 균일최강력 검정법

어떤 검정방법이 복합대립가설하에서의 모든 가능한 모수의 값에 대하여 최강력 검정이면 그 검정을 **균일최강력검정**이라 하고 기각역 C^* 을 크기 α 인 **균일최강력기각역**이라 부른다.

$\Rightarrow$ 특히 최량기각역이 대립가설 하에서의 모수의 값 θ_1 에 의존하지 않는 경우 즉, $H_1 : \theta = \theta_1 \,(> \theta_0)$ 에 대한 최강력 검정법이 $H_1 : \theta > \theta_0$ 에서도 최강력검정법이 된다는 말이다.

ex 8)에서 기각영역 $C = \left\{ (X_1, \cdots, X_n) \,\middle|\, \overline{X} \geq \mu_0 + z_\alpha \frac{1}{\sqrt{n}} \right\}$ 는 대립가설로 임의의 $\mu(> \mu_0)$ 을 택하여도 최량기각역이고 $P\left(\overline{X} \geq \mu_0 + z_\alpha \frac{1}{\sqrt{n}} \,\middle|\, \mu_0 \right) = \alpha$ 를 만족한다.

따라서, $H_0 : \mu = \mu_0$, $H_1 : \mu > \mu_0$ 에 대한 유의수준 α 인 균일 최강력 기각역이 된다.

또 $\pi(\mu)$ 가 μ 에 대한 증가함수이므로 $\max\{\pi(\mu) \mid \mu \leq \mu_0\} = \pi(\mu_0)$ 이므로 $H_0 : \mu \leq \mu_0$ 와 $H_1 : \mu > \mu_0$ 에 대한 균일최강력기각역이 될 수도 있음을 알 수 있다.

Def : 단조우도비

확률표본 $X_1, X_2, \cdots, X_n$ 의 결합확률밀도함수가 $f(x_1, \cdots, x_n ; \theta)$ 일 때 우도비 $\frac{L(\theta_1)}{L(\theta_2)}$ 가 $\theta_1 < \theta_2$ 에 대하여 $t(x_1, x_2, \cdots, x_n)$ 의 비감소함수 또는 비증가함수이면 통계량 $T(X_1, X_2, \cdots, X_n)$ 은 **단조우도비**의 성질이 있다고 한다.

ex 9)

$X_1, X_2, \cdots, X_n$ 이 지수분포에서 얻은 확률표본이다.

$\lambda_1 < \lambda_2$ 에 대한 우도비 $\frac{L(\lambda_1)}{L(\lambda_2)} = \left(\frac{\lambda_2}{\lambda_1}\right)^n \exp\left[-\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) \sum_{i=1}^{n} x_i \right]$ 일 때 단조우도비의 성질을 가지는 통계량 $T(X_1, X_2, \cdots, X_n)$ 를 구하여라.

Thm : 단조우도비를 이용한 균일최강력 검정법

결합확률밀도함수 $f(x_1,\ x_2,\ \cdots\ ,\ x_n\,;\theta)$가 통계량 $T(X_1,\ \cdots\ ,\ X_n)$에 대해 비증가하는 단조우도비를 갖는다면

(1) $H_0\ :\ \theta \le \theta_0$ 와 $H_1\ :\ \theta > \theta_0$ 에 대한 유의수준 α인

균일최강력 기각역은 $C=\{(X_1,\ \cdots\ ,\ X_n)\ :\ T(X_1,\ \cdots\ ,\ X_n) \ge k\}$ 이며

여기서 k는 $P[\,T(X_1,\ \cdots\ ,\ X_n) \ge k \mid \theta_0] = \alpha$에 의해 결정된다.

(2) $H_0\ :\ \theta \ge \theta_0$ 와 $H_1\ :\ \theta < \theta_0$ 에 대한 유의수준 α인

균일최강력 기각역은 $C=\{(X_1,\ \cdots\ ,\ X_n)\ :\ T(X_1,\ \cdots\ ,\ X_n) \le k\}$ 이며

여기서 k는 $P[\,T(X_1,\ \cdots\ ,\ X_n) \le k \mid \theta_0] = \alpha$에 의해 결정된다.

ex 10)

$X_1,\ X_2,\ \cdots\ ,\ X_n$은 정규분포 $N(\mu,\ 1)$에서 얻은 확률표본이다.

$H_0\ :\ \mu \le \mu_0$와 $H_1\ :\ \mu > \mu_0$를 검정하기 위한 균일최강력기각역을 구하여라.

ex 11)

$X_1,\ X_2,\ \cdots\ ,\ X_n$이 지수분포에서 얻은 확률표본이라 할 때

$H_0\ :\ \lambda \le \lambda_0$와 $H_1\ :\ \lambda > \lambda_0$를 검정하기 위한 균일최강력기각역을 구하여라.

제 10 장　범주형 자료

Def (54) : 다항분포(multinomial distribution)

한 번의 시행결과 나올 수 있는 것이 A_1, A_2, $\cdots$, A_k $(k \geq 3)$ 중의 하나이고 각각의 사상이 출현할 확률이 각각 P_1, P_2, $\cdots$, P_k이며 $P_1 + P_2 + \cdots + P_k = 1$이라 하자.
이 때, n회 시행하여 그 중에 A_1, A_2, $\cdots$, A_k가 출현하는 횟수를 x_1, x_2, $\cdots$, x_k라 하고 $x_1 + x_2 + \cdots + x_k = n$에 관한 확률분포를 **다항분포**라 한다.

$$\Rightarrow \quad P(X_1 = x_1,\ X_2 = x_2,\ \cdots,\ X_k = x_k) = \frac{n!}{x_1!\,x_2!\,\cdots\,x_k!} P_1^{x_1} P_2^{x_2} \cdots P_k^{x_k} \qquad (k \geq 3)$$

예제 331
주사위를 10번 던질 때 1의 눈이 1번, 2의 눈이 1번, 3의 눈이 2번, 4의 눈이 2번, 5의 눈이 2번, 6의 눈이 2번 나올 확률을 구하여라.
- 풀이 -
(판서 중 오타있습니다. 6!이 아니고 10! 입니다. 아래 풀이를 참고하세요)

$$P(X_1 = 1,\ X_2 = 1,\ X_3 = 2,\ X_4 = 2,\ X_5 = 2,\ X_6 = 2)$$
$$= \frac{10!}{1!\,1!\,2!\,2!\,2!\,2!} \left(\frac{1}{6}\right)\left(\frac{1}{6}\right)\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2\left(\frac{1}{6}\right)^2$$
$$= 0.003751$$

Thm (65): 다항분포의 검정(카이제곱 검정)

k개의 범주를 갖는 다항분포에서 i번째 범주의 속성확률을 p_i라 할 때

귀무가설 H_0 : $p_1 = p_{10}$, $p_2 = p_{20}$, $\cdots$, $p_k = p_{k0}$

대립가설 H_1 : 적어도 하나의 i에 대하여 $p_i \neq p_{i0}$일 때

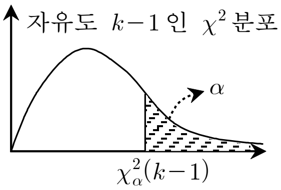
① 검정통계량

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}}$$

(χ^2은 근사적으로 자유도 $k-1$인 카이제곱분포를 따른다.)

② 유의수준 α인 기각역

$$\chi^2 \geq \chi^2_\alpha(k-1)$$



⇒ [각 셀의 관측값과 기댓값]

범주	$i=1$	$i=2$	$\cdots$	$i=k$	행합계
관측값	n_1	n_2	$\cdots$	n_k	n
기댓값	nP_{10}	nP_{20}	$\cdots$	nP_{k0}	n

㉠ 주사위 120번

범주	1	2	3	4	5	6	합
관측값	18	18	22	20	21	21	120
기댓값	20	20	20	20	20	20	120

예제 332

주사위를 60번 던졌을 때 나온 눈의 횟수가 다음 표와 같을 때 주어진 자료에 근거하여 주사위가 공정한가를 유의수준 $\alpha = 0.05$에서 검정하여라.

눈	1	2	3	4	5	6
관찰횟수	12	10	9	8	13	8
기대횟수	10	10	10	10	10	10

Thm (66) : 적합도 검정(goodness-of-fit test)

관측값과 기댓값 사이의 적합도 검정은

$H_0 : p_1 = p_{10},\ p_2 = p_{20},\ \cdots,\ p_k = p_{k0},\quad H_1 : H_0$가 아님

으로 주어졌을 때

① 검정통계량 $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - E_i)^2}{E_i}$, f_i는 관측값, E_i는 기댓값

② 유의수준 α인 기각역

$\chi^2 \ge \chi^2_\alpha(k-1)$

예제 333

멘델의 유전법칙은 동형접합으로 발생할 확률이 $\frac{9}{16}$, $\frac{3}{16}$, $\frac{3}{16}$, $\frac{1}{16}$이 되는 것이다. 큰 cut-leaf 토마토와 작은 potato-leaf 토마토가 동형접합시킨 후 다음과 같은 자료를 얻었다. 이 자료에 근거하여 토마토의 동형접합이 멘델의 법칙을 따른다고 할 수 있는지를 유의수준 $\alpha = 0.10$에서 검정하여라.

	$i=1$	$i=2$	$i=3$	$i=4$
n_i	926	288	293	104
np_{i0}	906.2	302.1	302.1	100.7
	Tall cut-leaf	Tall potato-leaf	Dwarf cut-leaf	Dwarf potato-leaf

예제 334

사람들은 0에서 9까지의 정수열을 무작위로 적는 경우에 연속으로 같은 수를 적을 확률은 $\frac{1}{10}$, 차이가 1인 수를 적을 확률은 $\frac{2}{10}$(여기서 0은 9와 1차이로 가정함), 그 외는 $\frac{7}{10}$이라고 한다. 다음 자료를 이용하여 아래 가설을 검정하여라. (단, 유의수준 $\alpha = 0.05$) $H_0 : p_1 = \frac{1}{10},\ p_2 = \frac{2}{10},\ p_3 = \frac{7}{10}$, $H_1 : H_0$가 아님

※어떤 사람이 50개의 정수를 적은 결과

	연속같은 수	차이가 1인 수	그 외의 수
관측도수	0	8	42
기대도수	5	10	35

예제 335

균형인 동전 4개를 무작위로 던지는 실험에서 X를 앞면이 나온 동전의 수라 하자. 매 실험이 독립적으로 시행된다고 할 때 X는 이항분포 $B\left(4,\ \frac{1}{2}\right)$를 가진다. 이 실험을 100번 반복시행 했을 때 앞면이 나온 동전수가 0, 1, 2, 3, 4인 도수가 각각 7, 18, 40, 31, 4번이었다. 이 결과는 X의 분포에 대한 가정을 뒷받침하는가를 유의수준 $\alpha = 0.05$로 검정하여라.

예제 336

유전법칙에 의하여 어떤 식물의 품종이 4종류로 나타나는데 그 비는 9 : 3 : 3 : 1 이라고 한다. 실험에서 4가지 품종이 각각 150, 80, 70, 20으로 나타났다면 이 결과는 유전법칙에 부합되는지 검정하라. (단, 유의수준 $\alpha=0.05$)

예제 337

시간이 지남에 따라 일어난 어떤 특정사건의 발생횟수가 다음 표와 같이 주어졌을 때 이 자료가 특정사건의 발생횟수가 포아송 분포를 따르는지를 검정하여라.
(단, 유의수준 $\alpha=0.05$, r개의 모수가 미지일 때 자유도는 $k-1-r$로 둔다.)

발생건수	0	1	2	3	4	5	6	7	8	9
관찰도수	6	20	36	50	52	40	25	12	4	5
기대도수	5.4	20.6	39.5	50.7	48.7	37.4	24.0	13.2	6.3	9.6

예제 338

다음 자료는 119에 걸려온 전화들을 시간간격으로 정리한 것이다. 이 자료는 시간간격 X가 평균 $\theta = 20$인 지수분포를 따르는지를 검정하여라.
(단, 유의수준 $\alpha = 0.05$)

구간	도수	확률	기대도수
$A_1 = [0,\ 9]$	41	0.362	38.0520
$A_2 = (9,\ 18]$	22	0.231	24.2655
$A_3 = (18,\ 27]$	11	0.147	15.4665
$A_4 = (27,\ 36]$	10	0.094	9.8700
$A_5 = (36,\ 45]$	9	0.060	6.2895
$A_6 = (45,\ 54]$	5	0.038	4.0110
$A_7 = (54,\ 63]$	2	0.024	2.5620
$A_8 = (63,\ 72]$	3	0.016	1.6275
$A_9 = (72,\ \infty)$	2	0.027	2.8665

Thm (67) : 독립성 검정

두 속성간에 독립관계의 여부를 판단하는 검정이다.

귀무가설 H_0 : $p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$ $(i=1,\ 2,\ \cdots,\ l,\ j=1,\ 2,\ \cdots,\ m)$

대립가설 H_1 : H_0가 아니다.

검정통계량 $\chi^2 = \sum\sum \frac{\left(f_{ij} - \widehat{E_{ij}}\right)^2}{\widehat{E_{ij}}}$ (단, f_{ij}는 관측도수, E_{ij}는 도수의 기대치)

은 자유도 $(l-1)(m-1)$인 카이제곱분포를 따른다.

유의수준 α인 기각역은 $\chi^2 \ge \chi^2_{\alpha}((l-1)(m-1))$ 이다.

$\Rightarrow$ $l \times m$ 분할표

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_m	계
A_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1m}	$T_{1\bullet}$
A_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2m}	$T_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{im}	$T_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_l	f_{l1}	f_{l2}	$\cdots$	f_{lj}	$\cdots$	f_{lm}	$T_{l\bullet}$
계	$T_{\bullet 1}$	$T_{\bullet 2}$	$\cdots$	$T_{\bullet j}$	$\cdots$	$T_{\bullet m}$	T

예제 339

다음 자료는 어떤 도시의 고등학교 학군과 학력고사 성적에 대한 3×5 분할 표이다. 이 자료에 근거하여 학군과 성적 사이의 독립여부를 유의수준 5%로 검정하여라.

성적 \ 학군	1	2	3	4	5	합
1등급	23	33	50	20	24	150
2등급	18	35	45	39	63	200
3등급	19	37	45	21	28	150
합	60	105	140	80	115	500

예제 340

다음과 같이 3×4 분할표에서 작업조와 기계호기간에 고장횟수의 발생은 서로 독립적인지를 유의수준 $\alpha = 0.05$ 에서 검정하여라.

작업조 \ 기계별	1 호기	2 호기	3 호기	4 호기	작업조 계
제 1조	6	13	10	12	41
제 2조	12	21	10	19	62
제 3조	10	18	13	13	54
기계별 계	28	52	33	44	157

예제 341

다음 2×5 분할표에서 교육방법과 성적 사이의 독립여부를 $\alpha=0.05$ 에서 검정하라.

등급	1	2	3	4	5	합계
방법 I	8	13	16	10	3	50
방법 II	4	9	14	16	7	50
합계	12	22	30	26	10	100

Thm (68) : 동질성 검정(=동일성 검정)

r개의 모집단에 대한 동질성에 대한 통계적 가설검정이다.

H_0 : $p_{j|i} = p_j$ $(i=1, \cdots, r, \ j=1, \cdots, m)$ H_1 : H_0가 아니다.

검정통계량 $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{m}\frac{(f_{ij}-\widehat{E_{ij}})^2}{\widehat{E_{ij}}}$ (자유도 $=(r-1)(m-1)$)

유의수준 α인 기각역

$\chi^2 \geq \chi^2_\alpha((r-1)(m-1))$

$\Rightarrow$

예제 342

흙에서 사는 네 종류의 벌레에 대한 생태를 조사하기 위하여 모래, 찰흙, 진흙의 세 가지 흙에서 각각 100마리의 벌레를 채취한 결과가 다음 표와 같다면 흙의 종류에 따라 벌레의 분포가 모두 같은가를 $\alpha = 0.05$에서 검정하여라.

	종류 1	종류 2	종류 3	종류 4	표본크기
모래	27	24	23	26	100
찰흙	20	32	18	30	100
진흙	13	37	16	34	100
합	60	93	57	90	300

예제 343

성별에 따른 자동차 모델의 선호도를 조사한 것이 아래의 표이다.

남자와 여자의 성별에 따라 선호도가 같은지를 유의수준 $\alpha = 0.05$ 에서 검정하여라.

성별 \ 모델	A	B	C	D	표본크기
남자	120	89	173	118	500
여자	124	155	147	74	500
합	244	244	320	192	1000

제 11 장 상관과 회귀분석

Def (55) : 상관분석(correlation analysis)

두 확률변수 간에 상호 관련성의 강도에 관한 통계적 분석방법을 **상관분석**이라고 한다. 이러한 관련성의 정도의 측도를 **상관계수**라 한다.

㉮ 훈련시간과 생산성, 오존의 농도와 탄소의 농도 등

Def (56) : 상관계수(correlation coefficient)

두 변수간의 상관관계를 측정하는 측도를 **상관계수**라 한다.

① 모상관계수 $\rho = \dfrac{Cov(X,\ Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ $(-1 \le \rho \le 1)$

② 표본상관계수 $r = \dfrac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \dfrac{\sum x_i y_i - \dfrac{\sum x_i \sum y_i}{n}}{\sqrt{\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}}\sqrt{\sum y_i^2 - \dfrac{\left(\sum y_i\right)^2}{n}}}$

①⇒

②⇒

예제 344

다음 표본 자료에서 두 변수 X와 Y에 대한 표본상관계수와 공분산을 구하여라.

X	1	2	3	4	5	6	7
Y	6	7	9	13	11	15	16

예제 345

다음 표본자료에서 훈련시간과 생산성에 대한 표본상관계수를 구하여라.

훈련시간(X)	10	20	30	40	50	60	70
생산성(Y)	40	45	50	65	70	70	80

예제 346

8개의 서로 다른 토양에 대한 감자와 밀의 산출량이 다음 표본 자료로 주어졌을 때 감자와 밀의 산출량에 대한 표본상관계수 r_{xy}를 구하여라.

감자(X)	2.4	3.4	4.6	3.7	2.2	3.3	4.0	2.1
밀(Y)	1.33	2.12	1.80	1.65	2.00	1.76	2.11	1.63

예제 347

다음은 오존의 농도와 탄소의 농도 표본자료이다.

이들의 표본상관계수 r_{xy}를 구하여라.

오존의 농도(X)	0.066	0.088	0.120	0.050	0.162	0.186	0.057	0.100	0.112	0.055	0.154	0.074	0.111	0.140	0.071	0.110
탄소의 농도(Y)	4.6	11.6	9.5	6.3	13.8	15.4	2.5	11.8	8.0	7.0	20.6	16.6	9.2	17.9	2.8	13.0

〈상관관계의 성질〉

① 상관계수의 범위 : $-1 \le r \le 1$

② $r = -1$: 음의 상관관계 (또는 역상관)

③ $r = 0$: 무상관관계 (X, Y는 독립적)

④ $r = 1$: 양의 상관관계 (또는 정상관)

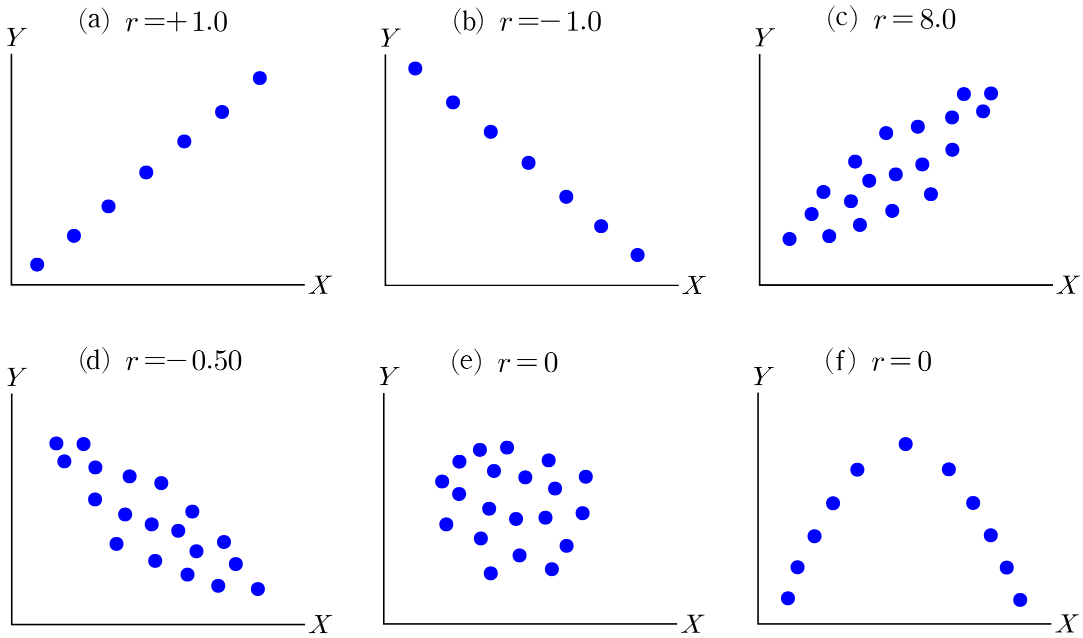
⑤ X, Y의 상관관계가 r일 때

$aX+b$, $cY+d$의 상관계수는 $ac > 0$이면 r이고, $ac < 0$이면 $-r$이다.

〈상관계수의 해석〉

상관계수의 범위	상관관계의 언어적 표현
0.00 ~ 0.20	상관이 거의 없다.
0.20 ~ 0.40	상관이 낮다.
0.40 ~ 0.60	상관이 있다.
0.60 ~ 0.80	상관이 높다.
0.80 ~ 1.00	상관이 매우 높다.

〈산포도와 상관계수의 모양〉



Thm (69): 상관계수의 성질

두 확률변수 X와 Y사이의 상관계수 $\rho(X,\ Y)$는

① $\rho(X,\ Y) = \rho(Y,\ X)$

② $\rho(X,\ Y) = \rho(aX+b,\ cY+d)$

(단, a와 c는 0이 아닌 같은 부호이며 b와 d는 임의의 상수)

③ $|\rho(X,\ Y)| \le 1$

④ $|\rho(X,\ Y)| = 1$의 필요충분조건은 $Y = aX+b$

(단, $a \ne 0$, b는 임의의 상수)

⑤ 확률변수 X와 Y가 서로 독립이면 $\rho(X,\ Y) = 0$이다.

(단, 역은 성립하지 않는다.)

$\Rightarrow$

예제 348

확률변수 X, Y의 결합확률밀도함수가

$f(x,\ y)=\begin{cases} x+y, & (0<x,\ y<1) \\ 0, & (\text{기타}) \end{cases}$ 인 경우일 때 X와 Y의 상관계수를 구하여라.

예제 349

다음 표본자료에서 훈련시간과 생산성에 대한 상관계수를 구하여라.

훈련시간(X)	10	20	30	40	50	60	70
생산성(Y)	40	45	50	65	70	70	80

예제 350

다음 표본자료에서 확률변수 X와 Y에 대한 상관계수를 구하여라.

X	320	356	368	341	335	321	348	382	376	340	358	325
Y	2.8	3.1	3.3	3.0	2.9	3.1	3.6	3.8	3.5	2.2	3.9	3.2

Thm (70) : 결정계수 r^2

결정계수는 예측변수(X)가 종속변수(Y)를 설명해주는 비율을 말한다.

※ 결정계수$\left(r^2\right)$와 Y분산 $\left(S_Y^2\right)$간의 관계

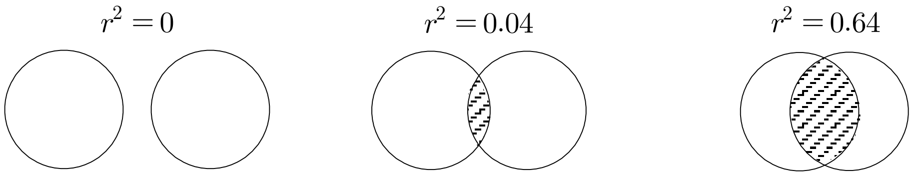
Y에 대한 X의 설명부분$= r_{XY}^2 \cdot S_Y^2$

Y에 대한 X의 설명되지 않는 부분$= \left(1 - r_{XY}^2\right) S_Y^2$

$\Rightarrow$ 만약 X와 Y를 표준화시키면 $S_Y^2 = 1$ 이므로

Y에 대한 X의 설명부분$= r_{XY}^2$

Y에 대한 X의 설명되지 않는 부분$= 1 - r_{XY}^2$



Thm (71) : 이관계수 k(coefficient of alienation)

한 변수로 다른 변수를 예측할 때 생기는 오차의 정도를 **이관계수**라 한다. 이관계수 $k = \sqrt{1 - r^2}$ 으로 계산된다.

예제 351

다음 표본자료에서 확률변수 X와 Y에 대한 상관계수, 결정계수, 이관계수를 구하여라.

미적분학 점수(X)	10	20	30	40	50	60	70
통계학 점수(Y)	40	45	50	65	70	70	80

(풀이)

예제 349에서 $r_{XY} = 0.98$

$$r_{XY}^2 = 0.9604$$

$$k = \sqrt{1 - 0.9604} = 0.199$$

Thm (72): 모상관계수 ρ에 대한 검정

① $H_0 : \rho = 0, \quad H_1 : \rho \neq 0$

② 검정통계량 $T = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ (자유도 $n-2$인 t분포를 따른다.)

③ 유의수준 α에 따른 기각역

$|T| \geq t_{\frac{\alpha}{2}}(n-2)$

예제 352

어느 회사에서 근무연수(X)와 직장만족도(Y)의 상관계수를 구했더니 $r = 0.457$이 나왔다. 상관계수 r의 유의도 검정을 $\alpha = 0.01$에서 하라. $(n = 10)$

예제 353

다음 표본자료에서 확률변수 X와 Y의 상관유무를 유의수준 $\alpha = 0.05$에서 검정하여라.

X	79	78	83	86	85	87	90	94
Y	84	70	88	75	81	75	88	90

예제 354

어느 대학 통계학과 학생들의 미적분학 점수와 통계학 점수의 상관관계를 조사하려고 10명을 임의추출하여 조사한 표본자료이다. 두 과목의 점수의 상관유무를 유의수준 $\alpha = 0.05$에서 검정하여라.

미적분학 점수(X)	82	75	95	99	85	70	77	60	63	64
통계학 점수(Y)	79	80	90	89	91	65	67	62	61	66

Thm (73) : 모상관계수 ρ의 구간추정

$\rho \neq 0$ 에서 얻은 표본의 크기 n이 충분히 클 때

r을 $z = \frac{1}{2}\ln\frac{1+r}{1-r} = \tanh^{-1} r$ 으로 변환하면 z는

근사적으로 $N\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho},\ \frac{1}{n-3}\right)$를 따른다.

$$\frac{1}{2}\ln\frac{1+r}{1-r} - z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}} \leq \frac{1}{2}\ln\frac{1+\rho}{1-\rho} \leq \frac{1}{2}\ln\frac{1+r}{1-r} + z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$$

에서 ρ의 구간추정을 구할 수 있다.

예제 355

임의로 추출한 표본크기 $n=10$ 에서 구한 표본상관계수는 $r=0.89$ 이다. 이 때, 모상관계수 ρ의 95% 신뢰구간을 구하여라. (단, 정규분포 근사법을 이용하라.)

예제 356

다음 표본자료는 기억력(X)와 판단력(Y)에 관한 데이터이다.

기억력(X)	11	10	14	18	10	5	12	7	15	16
판단력(Y)	6	4	6	9	3	2	8	3	9	7

(1) 기억력과 판단력의 상관계수를 구하여라.

(2) H_0 : $\rho = 0$, H_1 : $\rho \neq 0$ 를 유의수준 $\alpha = 0.05$ 에서 검정하여라.

(3) 모상관계수 ρ 의 신뢰율 95% 구간을 추정하라.
(단, 정규분포근사법을 이용하라.)

Def (57) : 회귀분석(regression analysis)

회귀분석이란 둘 또는 그 이상의 변수들 간의 인과관계를 파악함으로써 어떤 특정한 변수(종속변수)의 값을 다른 한 개 또는 그 이상의 변수(독립변수)들로부터 설명하고 예측하는 추측통계의 한 분야이다.

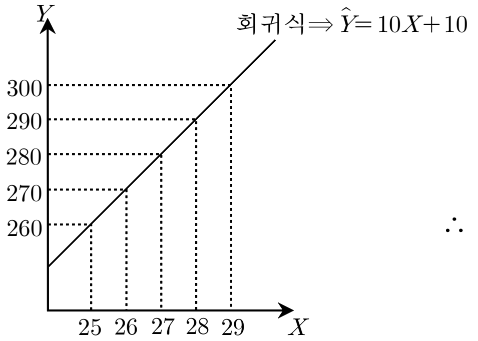
$\begin{cases} \text{단순회귀분석} \Rightarrow \text{독립변수가 1개} \\ \text{다중회귀분석} \Rightarrow \text{독립변수가 2개 이상} \end{cases}$

※ 회귀식 ⇒ 특정변수값의 변화와 다른 변수값의 변화가 가지는 수학적 선형함수식을 파악함으로써 상호관계를 추론하게 되는데 추정된 함수식을 회귀식이라 한다. 회귀식이 타당시 되는 경우 독립변수의 값을 기초로 종속변수의 값을 예측할 수 있다.

예 최고 기온과 에어컨 판매(A대리점에서)

온도(X)	25	26	27	28	29
에어컨 판매대수(Y)	260	270	280	290	300

"최고 기온이 32도 일 때 판매량은 얼마나 예측되는가"
이것을 알아보는 게 회귀분석의 목적이다.

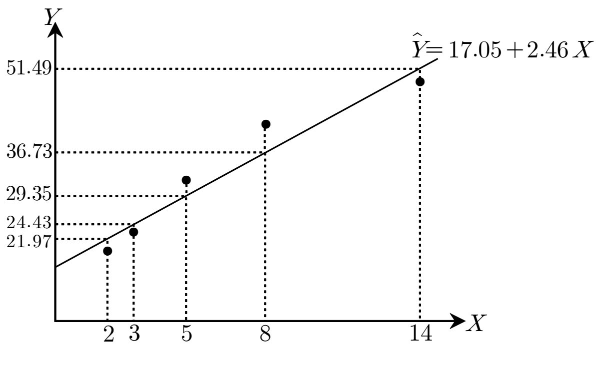


$\therefore\ x=32$ 일 때 $\hat{Y}=(10)(32)+10=330$ 대

예

X	2	3	5	8	14
Y	20	24	31	39	50

⇒ 회귀식 $\hat{Y}=17.05+2.46X$



Def (58): 단순선형회귀모형

오차항 ϵ을 수반하는 모집단의 **단순선형회귀모**형은 다음과 같다.

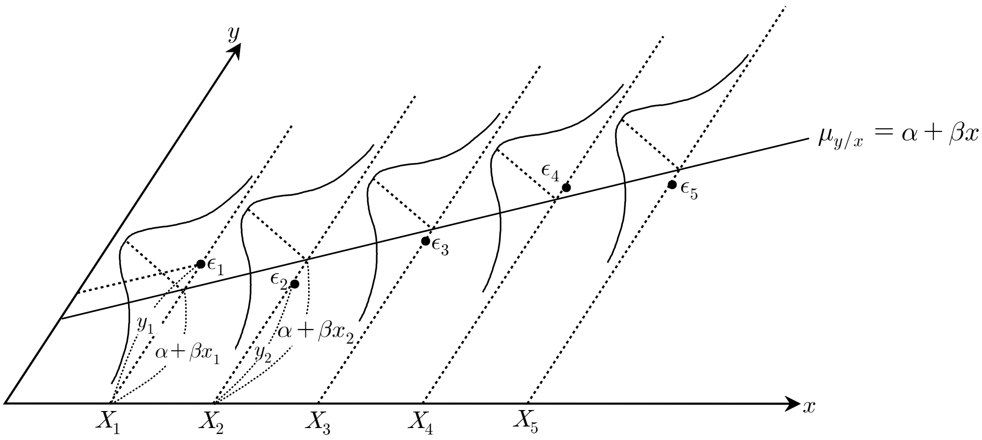
$$y = \alpha + \beta x + \epsilon$$

α와 β는 각각 미지의 절편과 기울기 모수이며 회귀변수 x는 고정된 수학변수로 가정한다.

랜덤오차 ϵ은 $E(\epsilon) = 0$이고 $\mathrm{Var}(\epsilon) = \sigma^2$인 정규분포를 따르고 ϵ_i와 ϵ_j는 서로 독립이다. 또, ϵ_i는 모든 i에 대해 동일분산을 갖는다.

① $E(\epsilon) = 0$이므로 $E(y) = \alpha + \beta x$, $V(y) = V(\epsilon) = \sigma^2$이다.
따라서, y는 확률변수이고 $N(\alpha + \beta x, \sigma^2)$을 따른다.
여기서, $E(y) = \mu_{y/x} = \alpha + \beta x$를 모집단 회귀식(=모회귀식) 또는 모회귀직선이라고 부른다.

② $\epsilon = y - (\alpha + \beta x)$



③ 모회귀계수 α, β를 구하기 위해 모집단 전체를 조사하는 것은 불가능하다. (비용과 시간) 그래서 이들은 표본자료를 이용하여 추정된다.

$\mu_{y/x} = \alpha + \beta x$ (모회귀선)

$\hat{y} = a + bx$ (표본회귀선 or 추정회귀선)

Def (59) : 잔차(residual)

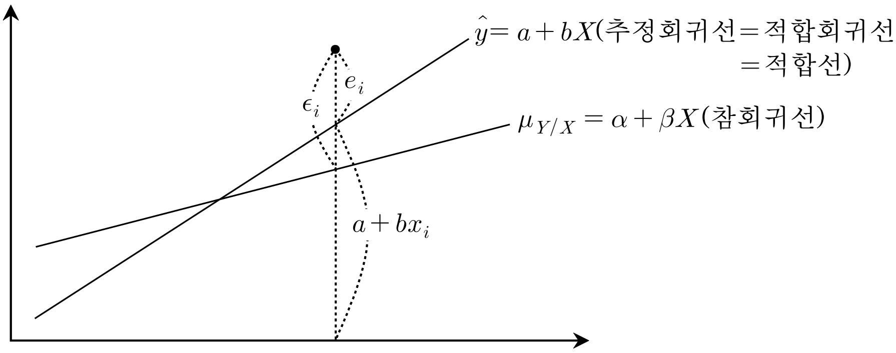
잔차란 적합모형 $\hat{y}=a+bx$를 적합시킬 때 생기는 오차이다.

회귀자료 $\left[(x_i,\ y_i)\,;\, i=1,\ 2,\ \cdots\ ,\ n\right]$와 적합모형 $\hat{y_i}=a+bx_i$가 주어졌을 때 i번째 잔차 $e_i=y_i-\hat{y_i}$ $(i=1,\ 2,\ \cdots\ ,\ n)$이다.

여기서, a와 b는 회귀계수 α와 β의 추정치이다.

⇒ 잔차가 크면 모형의 적합성은 좋지 않다고 할 수 있고,
반대로 잔차가 작으면 적합성이 좋다고 할 수 있다.

※ 모형오차 ϵ_i(실제 관측될 수 없다)와 잔차 e_i(관측되고, 분석과정에서 매우 중요)



⇒ $y_i=a+bx_i+e_i$임을 알 수 있다.

∴ 적합선으로부터 예측치 $(\hat{y})$를 구할 수 있고 선형관계의 강도와 적합모형의 적절성에 관한 정보를 얻을 수 있다.

실제 회귀선을 추정하려면 즉, α의 추정치 a, β의 추정치 b는 최소제곱추정법으로 구한다.

(오차 e_i를 가장 작게 하는 회귀계수 a와 b를 결정하는 것이 합리적이다.
여기서 오차 e_i를 개별적으로 최소화 할 수 없기 때문에
오차들의 합인 $\sum e_i$를 가능한 0에 가깝게 만들어야 한다.)

Thm (74) : 회귀계수의 추정(최소제곱법)

회귀계수 α와 β의 최소제곱추정값 a와 b는 다음과 같이 계산된다.

$$a = \frac{\sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i}{n} = \overline{y} - b\overline{x}$$

$$b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

※ 잔차 제곱합 (=회귀직선에 대한 오차 제곱합)는 기호로 SSE 로 나타낸다.

예제 357

표본자료 $(n=33)$에서 얻은 데이터가 다음과 같을 때 최소제곱법을 이용하여 회귀직선을 추정하여라.

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41355, \quad \sum_{i=1}^{33} x_i^2 = 41086$$

예제 358

다음 자료들을 최소제곱법에 의해 직선으로 적합시키고 공기속도가 190cm/sec일 때 연료방울의 기화계수값을 추정하라.

공기속도(cm/sec)x	20	60	100	140	180	220	260	300	340	380
기화계수(mm^2/sec) y	0.18	0.37	0.35	0.78	0.56	0.75	1.18	1.36	1.17	1.65

예제 359

어떤 실험에서 실험온도 x에서의 결과량 y가 다음 같은 데이터로 관찰되었다. 최소제곱법으로 회귀직선을 추정하여라.

x	30	35	40	45	50	55	60	65	70	75
y	25	32	34	43	42	48	55	56	72	70

예제 360

다음은 어떤 화학물의 첨가물의 양 (x)와 수율 (y)에 대한 자료이다. 최소제곱을 이용하여 회귀직선을 추정하여라.

x	1.7	2.3	2.8	3.5	4.2	4.9
y	33	35	50	45	66	63

예제 361

다음 자료는 어떤 화학물의 품질특성 (x)과 이 화학물을 원료로 써서 만든 제품의 특성 (y)에 관한 조사데이터이다.

x	36	40	34	44	33	36	40	33	26	36	31	44	41	29	42	35	45	31	32	38
y	29	32	29	40	31	29	34	30	25	31	28	34	34	27	33	33	34	27	30	33

(1) 상관계수 r를 구하여라.

(2) 모상관계수 ρ의 95% 신뢰구간을 구하여라.

(3) x와 y간의 회귀직선을 구하여라.

(4) 모상관계수 ρ가 $\rho=0$라는 귀무가설을 유의수준 $\alpha=0.05$에서 검정하여라.

예제 362
청바지를 판매하는 어떤 기업이 직영대리점 5곳을 대상으로 광고비와 판매량을 조사한 데이터가 다음과 같다.

광고비(X_i)	14	8	3	5	2
판매량(Y_i)	50	39	24	31	20

(단위: 백만원, 만벌이다.)

(1) $\widehat{Y_i}=a+bX_i$의 회귀식을 구하여라.

(2) 광고비를 백만원 더 지출하면 판매량은 얼마나 더 늘어나는가?

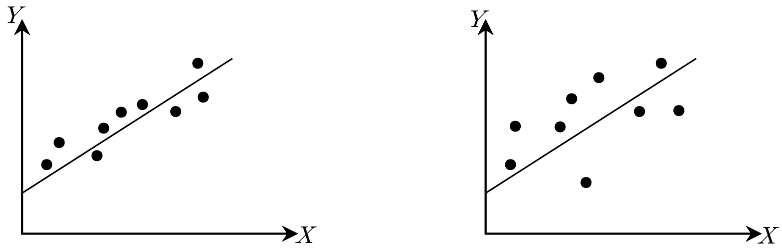
(3) 광고를 전혀 하지 않는다면 판매량은 얼마인가?

Def (60) : 표본회귀식에 대한 적합도 검정

표본회귀선이 두 변수간의 선형관계를 잘 나타내고 있는지를 알아보는 통계적 검정 방법을 **적합도 검정**이라 한다.

적합도 측정방법 { ① 추정의 표준오차를 가지고 평가하는 방법
② 결정계수를 가지고 평가하는 방법 }

※ 적합도가 다른 두 회귀식



Thm (75) : 추정의 표준오차에 의한 적합도 검정

$$S_e = \sqrt{\frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

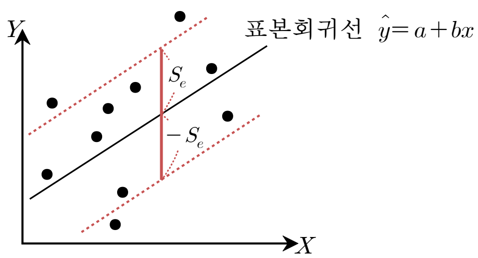
($\because$ S_e는 표본회귀식에서 얻은 예측치와 실제 관찰값 사이에서 나타나는 잔차들의 표준편차이다.)

① 관찰값의 68%가 회귀식의 $\pm S_e$ 안에 있다.

② 관찰값의 95%가 회귀식의 $\pm 2S_e$ 안에 있다.

③ 관찰값의 99.7%가 회귀식의 $\pm 3S_e$ 안에 있다.

※ 추정오차의 의미



예제 363
다음 표본자료를 이용하여 물음에 답하여라.

대상자	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
학력수준(X_i)	6	12	10	11	4	13	15	6	9	7	11	16	8	8	14
평점(Y_i)	4.0	3.5	3.0	2.5	4.5	2.5	1.0	5.5	3.5	1.5	3.0	2.0	2.0	3.0	3.5

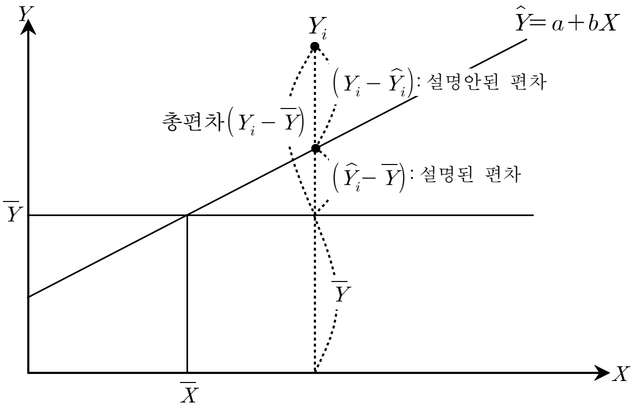
(1) 학력과 평점에 대한 표본회귀식을 구하여라.

(2) 모집단에 대한 회귀식의 추정오차 S_e를 구하여라.

Thm (76): 총편차

총편차=설명안 된 편차+설명된 편차

$\left(Y_i-\overline{Y}\right)=\left(Y_i-\widehat{Y}_i\right)+\left(\widehat{Y}_i-\overline{Y}\right)$



Thm (77): 총제곱합(SS_T)

총제곱합=잔차제곱합+회귀제곱합

$SS_T=SS_E+SS_R$

$\sum\left(Y_i-\overline{Y}\right)^2=\sum\left(Y_i-\widehat{Y}_i\right)^2+\sum\left(\widehat{Y}_i-\overline{Y}\right)^2$

※ 잔차의 성질 $\left(e_i=Y_i-\widehat{Y}_i\right)$

잔차합$=\sum e_i=\sum\left(Y_i-\widehat{Y}_i\right)=\sum\left(Y_i-a-bX_i\right)=0$

잔차들의 X에 대한 가중합$=\sum X_ie_i=\sum X_i\left(Y_i-\widehat{Y}_i\right)=\sum X_i\left(Y_i-a-bX_i\right)=0$

잔차들의 $\widehat{Y}$에 대한 가중합$=\sum \widehat{Y}_ie_i=\sum(a+bX_i)e_i=a\sum e_i+b\sum X_ie_i=0$

Thm (78): 결정계수에 의한 적합도 검정

결정계수에 의한 적합도 검정은 총제곱합 SS_T를 1이라 했을 때 회귀제곱합 SS_R이 차지하는 비율로 측정하는 방법이다.

$$\text{결정계수} = \frac{\text{설명된 제곱합 (회귀제곱합)}}{\text{총 제곱합}}$$

$$= 1 - \frac{\text{설명안 된 제곱합(잔차제곱합)}}{\text{총 제곱합}}$$

$$r^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

예제 364

예제 363 자료에서 결정계수 r^2을 구하여라.

예제 365

다음 자료를 이용하여 질문에 답하여라.

X_i	10	6	20	19	3	21	15	6	12	28
Y_i	9	4	18	11	5	16	13	10	17	20

(1) $\widehat{Y_i} = a + bX_i$ 의 회귀식을 구하여라.

이 자료에 의한 $\sum x_i$, $\sum y_i$, $\sum x_i y_i$, $\sum x_i^2$, $\sum y_i^2$ 값이 강의에서 잘못 나왔습니다. 다시 계산하여 아래와 같이 정리하였습니다.

- 풀이 -

$\sum x_i = 140$, $\sum y_i = 123$, $\sum x_i y_i = 2053$, $\sum x_i^2 = 2536$, $\sum y_i^2 = 1781$, $n = 10$

$$b = \frac{2053 - \dfrac{140 \times 123}{10}}{2536 - \dfrac{(140)^2}{10}} = \frac{331}{576} \fallingdotseq 0.5746$$

$$a = \frac{123 - 140 \times 0.5746}{10} = 4.2549$$

따라서, $\widehat{Y_i} = 4.2549 + 0.57X_i$ 입니다.

(2) SS_E 를 구하여라.

(3) 결정계수 r^2 을 구하여라.

예제 366

다음 자료는 P사의 주가(y)와 종합지수(x)에 관해 조사한 데이터를 정리한 것이다.

$\sum x_i = 1828$, $\sum y_i = 7660$, $\sum x_i y_i = 1400290$, $\sum x_i^2 = 334250$, $\sum y_i = 5868056$, $n = 10$

(1) P사 주가와 종합지수간의 표본상관계수 r을 구하여라.

(2) 회귀식을 구하여라.

(3) (1)의 결과를 이용하여 결정계수를 구하여라.

Thm (79) : 회귀계수 β의 추정량 $\hat{\beta}$의 평균과 분산

단순회귀모형 $Y_i = \alpha + \beta x_i + \epsilon_i$에서 β의 추정량 $\hat{\beta}$는

$\hat{\beta} = \dfrac{S_{(xy)}}{S_{(xx)}} = \dfrac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$ 이고, $E(\hat{\beta}) = \beta$, $Var(\hat{\beta}) = \dfrac{\sigma^2}{S_{(xx)}}$ 이다.

또, $\hat{\beta}$는 $N\left(\beta,\ \dfrac{\sigma^2}{S_{(xx)}}\right)$을 따른다.

$\Rightarrow$

※ 회귀모수 α의 추정량 $\hat{\alpha}$의 평균과 분산

$E(\hat{\alpha}) = \alpha$, $Var(\hat{\alpha}) = \dfrac{\sigma^2 \sum x_i^2}{n\sum(x_i - \overline{x})^2}$ 이고 $\hat{\alpha}$는 $N\left(\alpha,\ \dfrac{\sigma^2 \sum x_i^2}{n\sum(x_i - \overline{x})^2}\right)$을 따른다.

Thm (80)

$\hat{\beta} \sim N\left(\beta,\ \dfrac{\sigma^2}{S_{(xx)}}\right)$이고 $\dfrac{SSE}{\sigma^2}$는 자유도 $n-2$인 카이제곱분포를 따르며

$\hat{\beta}$과 $\dfrac{SSE}{\sigma^2}$은 서로 독립임이 알려져 있다.

이 때, $\dfrac{\hat{\beta}-\beta}{\sqrt{\dfrac{MSE}{S_{(xx)}}}}$는 자유도 $n-2$인 t분포를 따른다.

$\dfrac{\dfrac{\hat{\beta}-\beta}{\dfrac{\sigma}{\sqrt{S_{(xx)}}}}}{\sqrt{\dfrac{\dfrac{SSE}{\sigma^2}}{n-2}}}$는 t분포 정의에 의해 자유도 $n-2$인 t분포를 따른다.

Thm (81): β에 관한 추정

β의 $100(1-\alpha)\%$ 신뢰구간은

$$\hat{\beta}-t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{(xx)}}} \leq \beta \leq \hat{\beta}+t_{\frac{\alpha}{2}}\frac{S}{\sqrt{S_{(xx)}}} \quad \text{이다.}$$

여기서 $t_{\frac{\alpha}{2}}$는 자유도 $n-2$인 t분포의 값이다.

예제 367

예제 365에서 β의 95% 신뢰구간을 구하여라.

예제 368

다음 데이터를 이용하여 β의 95% 신뢰구간을 구하여라.

X	3	7	11	15	18	27	29	30	30	31	31	32	33	33	34	36
Y	5	11	21	16	16	28	27	25	35	30	40	32	34	32	34	37

36	36	37	38	39	39	39	40	41	42	42	43	44	45	46	47	50
38	34	36	38	37	36	45	39	41	40	44	37	44	46	46	49	51

Thm (82): β에 대한 가설검정

귀무가설 H_0 : $\beta = \beta_0$ 의 검정

검정통계량 $T = \dfrac{\hat{\beta} - \beta_0}{\sqrt{\dfrac{MSE}{S_{(xx)}}}}$ $\left(\because\ T\text{는 자유도 } n-2\text{인 } t\text{분포를 따른다.} \right)$

예제 369

예제 363에서 학력이 평점을 예측할 수 있는 회귀변수인지를 $\alpha = 0.05$ 에서 검정하여라.

예제 370

다음 자료를 이용하여 $H_0 : \beta = 0$, $H_1 : \beta \neq 0$를 $\alpha = 0.05$에서 검정하여라.

X	5.7	6.8	9.6	10.0	10.7	12.6	14.4	15.0	15.3	16.2	17.8	18.7	19.7	20.6	25.0
Y	119.0	121.3	118.2	124.0	112.3	114.1	112.2	115.1	111.3	107.2	108.9	107.8	111.0	106.2	105.0

예제 371

다음은 은하계에 대한 거리 (X)와 후퇴속도 (Y)에 관한 데이터이다.

선형회귀모형 $Y=\alpha+\beta x_i+\epsilon$, $\epsilon \sim N(0,\ \sigma^2)$에 대하여 회귀모수 β에 대한 가설 H_0 : $\beta=0$, H_1 : $\beta>0$을 $\alpha=0.05$에서 검정하여라.

거리(X)	22	68	108	137	255	315	390	405	685	700	1100
속도(Y)	750	2400	3200	4700	9300	12000	13400	14400	24500	26000	38000

Thm (83) : 단순회귀 모형에 대한 F검정

SSE(잔차제곱합)은 자유도 $n-2$, SSR(회귀제곱합)은 자유도 1인 카이제곱분포를 따른다고 할 때

H_0 : $\beta=0$, H_1 : $\beta\neq0$ 의 가설검정에서

점정통계량 $F=\dfrac{\dfrac{SSR}{1}}{\dfrac{SSE}{n-2}}=\dfrac{MSR}{MSE}$ 는 자유도 $(1,\ n-2)$인 F분포를 따른다.

※ 기각역의 경계값 $F_{\alpha}(1,\ n-2)$

예제 372

$\sum_{i=1}^{15}\left(\widehat{Y_i}-\overline{Y}\right)^2=5.767$, $\sum_{i=1}^{15}\left(Y_i-\widehat{Y_i}\right)^2=13.247$ 일 때 X, Y 두 변수 간에 회귀모형이 성립되는지를 $\alpha=0.05$ 에서 F검정하여라.

<단순회귀의 분산분석표>

분산원	제곱합	자유도	평균제곱	F 값
회귀	$SSR=\sum\left(\widehat{Y_i}-\overline{Y}\right)^2$	1	$MSR=\dfrac{SSR}{1}$	$\dfrac{MSR}{MSE}$
잔차	$SSE=\sum\left(Y_i-\widehat{Y_i}\right)^2$	$n-2$	$MSE=\dfrac{SSE}{n-2}$	

예제 373

다음 자료는 학력 (X)과 TV 시청시간 (Y)에 관한 자료이다.

학력(X)	5	7	8	8	9	10	10	12	15	16
시청시간(Y)	5	5	2	2	4	1	4	3	3	1

(1) 학력과 TV시청시간 사이에 회귀모형이 성립되는지를 $\alpha = 0.05$ 수준에서 $F-$검정하라.

(2) $F-$검정한 결과를 회귀분석표를 정리하라.

분산원	제곱합	자유도	평균제곱	F 값
회귀				
잔차				

예제 374

다음의 자료를 이용하여 물음에 답하여라.

x	1	1	2	3	4	4	5	6	6	7
y	2.1	2.5	3.1	3.0	3.8	3.2	4.3	3.9	4.4	4.8

(1) 단순회귀식을 구하여라.

(2) 회귀분석표를 만들고 유의수준 $\alpha = 0.05$ 에서 F 검정을 하여라.

(3) 결정계수 r^2 을 구하고 이로부터 상관계수 r 을 구하여라.

Thm (84) : 표본 중회귀식의 추정

표본회귀식 $\widehat{Y_i} = a + b_1 X_i + b_2 Z_i$ 에 대하여

$$b_1 = \frac{\sum x_i y_i \sum z_i^2 - \sum z_i y_i \sum x_i z_i}{\sum x_i^2 \sum z_i^2 - \left(\sum x_i z_i\right)^2} \qquad \text{단) } x_i = X_i - \overline{X}$$

$$z_i = Z_i - \overline{Z}$$

$$b_2 = \frac{\sum z_i y_i \sum x_i^2 - \sum x_i y_i \sum x_i z_i}{\sum x_i^2 \sum z_i^2 - \left(\sum x_i z_i\right)^2} \qquad y_i = Y_i - \overline{Y}$$

$a = \overline{Y} - b_1 \overline{X} - b_2 \overline{Z}$ 이다.

$\Rightarrow$ 단순회귀분석에와 마찬가지로 잔차제곱합 $\sum\left(Y_i - \widehat{Y_i}\right)^2 = \sum\left(Y_i - a - b_1 X_i - b_2 Z_i\right)^2$ 을 최소로 하는 b_1, b_2, a를 구한 값이다. ($\because$ 편미분으로)

예제 372

다음 자료에서 회귀식 $\widehat{Y_i} = a + b_1 X_i + b_2 Z_i$ 를 구하여라.

X_i	2	5	10	6	3	4	1	7	4	8
Z_i	1	2	3	3	1	5	2	4	4	5
Y_i	7	8	15	6	3	10	5	12	14	20

예제 373

예제 372에서 결정계수 R^2을 구하여라.

(풀이)

Y_i	7	8	15	6	3	10	5	12	14	20
$\widehat{Y_i}$	3.788	8.160	14.212	10.852	4.628	12.886	4.800	13.544	11.024	16.246
$\left(Y_i - \overline{Y}\right)^2$	9	4	25	16	49	0	25	4	16	100
$\left(Y_i - \widehat{Y_i}\right)^2$	10.317	0.026	0.621	23.542	2.650	8.329	0.040	2.384	8.857	14.093

Thm (85) : 중회귀분석에서의 F-검정

SSR이 자유도 k(독립변수의 개수), SSE가 자유도 $(n-k-1)$인 카이제곱분포를 따를 때 검정통계량

$F=\dfrac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}=\dfrac{MSR}{MSE}$ 는 자유도 $(k,\ n-k-1)$인 F분포를 따른다.

예제 374

예제 373에서 H_0 : $\beta_1=\beta_2=0$, H_1 : 두 회귀계수 중 적어도 하나는 0이 아니다을 유의수준 $\alpha=0.01$에서 검정하여라.

<중회귀 분석표>

분산원	제곱합	자유도	평균제곱	F 값
회귀	$SSR=\sum\left(\hat{Y_i}-\overline{Y}\right)^2$	k(독립변수의 개수)	$\dfrac{SSR}{k}=MSR$	$\dfrac{MSR}{MSE}$
잔차	$SSE=\sum\left(Y_i-\hat{Y_i}\right)^2$	$n-k-1$	$\dfrac{SSE}{n-k-1}=MSE$	

예제 375

다음은 종속변수 Y와 독립변수 X_1, X_2를 조사한 표이다.

X_{1i}	4	8	9	8	8	12	6	10	6	9
X_{2i}	4	10	8	5	10	15	8	13	5	12
Y_i	9	20	22	15	17	30	18	25	10	20

(1) 회귀식 $\widehat{Y_i}= a+b_1X_{1i}+b_2X_{2i}$를 구하여라.

(2) H_0 : $\beta_1=\beta_2=0$, H_1 : 두 모수 중 적어도 하나는 0이 아니다
를 유의수준 $\alpha=0.05$에서 검정하여라.

제 12 장　분산 분석

Def (61) : 분산분석(=anova)

분산분석이란 표본분산들을 분석함으로써 모집단의 동일성을 검정하는 통계적 기법이다.

일원분산분석 ⇨ 독립변수 하나, 이원분산분석 ⇨ 독립변수 두 개

※독립변수=요인(factor)

⇒ 분산분석은 독립변수를 몇 개의 수준 (or 범주)으로 나누고 각 수준에 따라 나누어진 집단 간의 평균차를 검정하는 것이다.

㈎ 광고효과→광고매채(독립변수) : 신문, TV, 라디오, 개별광고 - 일원분산분석

광고효과→ $\begin{cases} \text{광고매채(독립변수) : 신문, TV, 라디오, 개별광고} \\ \text{소비자 나이(독립변수) : 20대, 30대, 40대, 50대} \end{cases}$ - 이원분산분석

⇒ 두 집단의 평균 차이를 검정하기 위해서는 $t-$검정 or $Z-$검정을 하였지만
3개 이상의 집단의 평균의 차이를 검정할 때는 분산분석을 해야한다.

만약 3개의 집단에 대해서 $t-$검정을 이용한다면 $(\alpha=0.05)$

$H_0 : m_1 = m_2$, $H_0 : m_1 = m_3$, $H_0 : m_2 = m_3$으로 총 세 번의 검정을 해야하고 각각의 귀무가설들의 결과가 동시에 일어나니까

오류 $\alpha = 1-(1-0.05)^3 = 1-0.857 = 0.143$으로 오류가 일어날 확률은 14.3%로 엄청 높아진다.

이러한 부담을 없애면서 애초의 유의수준 α를 유지하고 여러 모집단의 평균차이를 검정하는 방법이 바로 분산분석이다.

⇒ 분산분석의 기본가정

① 모집단은 정규분포를 따른다.

② 모집단의 분산은 모두 같다.

③ 표본은 각 모집단에서 랜덤 추출한다.

④ 모집단 내에서의 오차나 모집단 간의 오차는 서로 독립적이다.

※ 처리=고려된 인자들의 수준끼리의 곱

물론, 인자가 한 개인 경우이면 인자의 수준이 곧 처리이다.

Thm (86) : 총제곱합(SST), 집단 간 제곱합(SSB), 집단내 제곱합(SSW)

$SST = \sum\sum\left(X_{ij} - \overline{X}\right)^2$ 여기서, $\overline{X}$는 전체평균

$SSB = \sum\sum\left(\overline{X_j} - \overline{X}\right)^2$ $\overline{X_j}$는 j번째 집단의 평균

$SSW = \sum\sum\left(X_{ij} - \overline{X_j}\right)^2$ X_{ij}는 j번째 집단의 i번째 관찰 값

$SST = SSB + SSW$ 이다.

〈일원분산분석의 자료구성〉

관찰번호 \ 집단	집단 1	집단 2	집단 3	$\cdots$	집단 j	
1	X_{11}	X_{12}	X_{13}	$\cdots$	X_{1j}	
2	X_{21}	X_{22}	X_{23}	$\cdots$	X_{2j}	
3	X_{31}	X_{32}	X_{33}	$\cdots$	X_{3j}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
i	X_{i1}	X_{i2}	X_{i3}	$\cdots$	X_{ij}	
	$\overline{X_1}$	$\overline{X_2}$	$\overline{X_3}$	$\cdots$	$\overline{X_j}$	$\overline{X}$

① 관찰값의 모형

$X_{ij} = \mu + \alpha_j + \epsilon_{ij}$

(μ : 전체평균, α_j : j번째 집단의 영향, ϵ_{ij} : j번째 집단에 있는 관찰값 i의 우연적 오차)

② 실제 통계값 $X_{ij} = \overline{X} + \left(\overline{X_j} - \overline{X}\right) + \left(X_{ij} - \overline{X_j}\right)$

예제 376

P 자동차 회사에서는 특정 모델의 승용차를 흰색, 은색, 검정색, 파란색의 네 가지 색상으로 생산, 판매하고 있다. 이 승용차의 색상이 차량 판매량과 관계가 있는지를 알아보기 위해 여섯 지역을 선택하여 각각의 판매량을 조사하였다. (단위 : 천 대)

색상 / 지역	흰색(1)	은색(2)	검정색(3)	파란색(4)	합계
1	26	31	27	30	
2	28	28	25	29	
3	25	30	28	32	
4	29	27	24	31	
5	27	29	26	32	
6	27	29	26	32	
평균	$\overline{X_1} = 27$	$\overline{X_2} = 29$	$\overline{X_3} = 26$	$\overline{X_4} = 31$	$\overline{X} = 28.25$

총제곱합(SST), 집단 간 제곱합(SSB), 집단내 제곱합(SSW) 를 각각 구하여라.

Thm (87) : 분산분석에서 $F-$통계량

SSB는 자유도 $J-1$, SSW는 자유도 $N-J$인 카이제곱분포를 따를 때 분산분석을 위한 검정통계량

$$F=\frac{MSB}{MSW}=\frac{\frac{SSB}{J-1}}{\frac{SSW}{N-J}}$$ 는 자유도 $(J-1,\ N-J)$인 F분포를 따른다.

예제 377

예제 376에서 H_0 : μ_1(흰색)$=\mu_2$(은색)$=\mu_3$(검정색)$=\mu_4$(파란색)

H_1 : 색상에 따라 판매량이 다를 수 있다

를 F검정하여라. (단, 유의수준 $\alpha=0.05$)

(풀이)

<분산분석표>

분산원	제곱합	자유도	평균제곱	F값
집단간				
집단내				

예제 378

다음 자료는 재료가 제품수출에 영향을 미치는지를 알아보기 위해 재료(A)를 네 가지 수준으로 하여 각 수준에서 동일하게 6번 반복하여 랜덤하게 행한 결과이다. 재료가 제품수출에 영향을 미치는지를 $\alpha = 0.05$ 수준에서 검정하라.

재료	A_1	A_2	A_3	A_4	합계
1	76	82	79	81	
2	80	75	87	74	
3	78	83	88	76	
4	79	78	86	78	
5	83	85	84	73	
6	74	80	82	70	
평균	78.33	80.50	84.33	75.33	$\overline{X} = 79.625$

(풀이)

<분산분석표>

분산원	제곱합	자유도	평균제곱	F값
집단간				
집단내				

Thm (88): Tukey 검정을 위한 HSD 임계값

$$HSD = q\sqrt{\frac{MSW}{n}}$$

(n은 집단 내 사례수, q는 α수준과 MSW의 df, 집단 수 k에 의한 통계값)

$\Rightarrow$ 분산분석에 의해 귀무가설이 기각된 경우 특정 집단 간의 차이가 생긴다.
이 때, 어느 특정집단간에 차이가 있는지를 검정하는 방법이다.
$\Rightarrow$ 특정 집단간의 평균의 차이가 HSD 보다 크면 α수준에서 유의하다고 본다.

예제 379
예제 377에서 색상의 차이가 판매량에 영향을 미친다고 하였다. 이 때, 네 종류의 색상 중에서 어느 것이 서로 차이가 나는지를 $\alpha = 0.05$에서 검정하라.

※ 독립변수의 설명력 η^2

$\eta^2 = \dfrac{SSB}{SST}$는 분산분석에서 독립변수에 의해 설명될 수 있는 분산과 총분산간의 비율을 나타내는 지수이다.

예제 380

아래 자료를 이용하여 다음 물음에 답하여라.

관찰수 \ 집단	집단 1	집단 2	집단 3
1	13.4	11.8	9.4
2	11.9	12.2	11.6
3	12.5	10.9	10.7
4	12.8	12.6	11.5
5	10.5	12.8	9.9
6	10.3	12.5	10.1
7	12.3	11.9	10.9

(1) 분산분석표(ANOVA Table)를 작성하고

H_0 : $\mu_1 = \mu_2 = \mu_3$, H_1 : 집단에 따라 평균이 다를 수 있다.

을 $\alpha = 0.05$ 의 수준에서 검정하여라.

(2) Tukey 방법에 의해 사후 비교를 하여라.

(3) 독립변수의 설명력 η^2(에타제곱)을 구하여라.

$$\eta^2 = \frac{SSB}{SST} = \frac{9.745}{24.433} = 0.399$$

∴ 독립변수인 각 집단은 종속변수에 대해 39.9%를 설명한다.

예제 381

다음 자료를 이용하여 분산분석표를 구하고 $H_0 : \mu_1 = \mu_2 = \mu_3$를 $\alpha = 0.05$ 수준에서 검정하라. 또 가능하다면 Tukey 방법에 의한 사후비교를 하여라.

관찰수 \ 집단	집단 1	집단 2	집단 3
1	28	24	15
2	22	17	22
3	21	20	20
4	29	19	19
	$\overline{X_1} = 25$	$\overline{X_2} = 20$	$\overline{X_3} = 19$

예제 382

다음 자료를 이용하여 질문에 답하여라.

관찰수 \ 집단	$N(\mu_1,\ \sigma^2)$	$N(\mu_2,\ \sigma^2)$	$N(\mu_3,\ \sigma^2)$	$N(\mu_4,\ \sigma^2)$
1	52	49	55	49
2	43	52	51	52
3	40	46	53	55
4	47	43	55	54
5	45	40	49	48
	$\overline{X_1}=45.4$	$\overline{X_2}=46$	$\overline{X_3}=52.6$	$\overline{X_4}=51.6$

(1) 분산분석표를 작성하고 H_0 : $\mu_1=\mu_2=\mu_3=\mu_4$ 를 $\alpha=0.05$ 수준에서 검정하여라.

(2) Tukey 방법으로 사후비교를 하여라.

Def (62) : 이원분산분석

이원분산분석은 두 개의 독립변수 중 어느 변수에 분산의 원인이 있는지 그리고 두 변수의 상호작용은 어떠한지를 알아보는 통계적 기법이다.

㉥ 광고효과→ { 요인 A(광고매체) : 인터넷, TV, 신문, 라디오
요인 B(소비자층) : 10대, 20대, 30대 }

<이원분산분석표>

분산원	제곱합	자유도	평균제곱	F값
인자 A (=A 효과)	SSA	$a-1$	$\frac{SSA}{a-1}=MSA$	$\frac{MSA}{MSW}$
인자 B (=B 효과)	SSB	$b-1$	$\frac{SSB}{b-1}=MSB$	$\frac{MSB}{MSW}$
교호작용 (=AB상호작용)	$SSAB$	$(a-1)(b-1)$	$\frac{SSAB}{(a-1)(b-1)}$ $=MSAB$	$\frac{MSAB}{MSW}$
오차 (=집단내)	$SSW(=SSE)$	$N-ab$ $=ab(k-1)$	$\frac{SSW}{ab(k-1)}$ $=MSE$	

※ $SST=SSA+SSB+SSAB+SSW$

① $SST=\sum\sum\sum X_{ijk}^2-\frac{T^2}{N}$ ($\because$ T=전체합, $N=abk$)

② $SSA=\sum_{i=1}^{a}\frac{T_{i\bullet\bullet}^2}{bk}-\frac{T^2}{N}$

③ $SSB=\sum_{j=1}^{b}\frac{T_{\bullet j\bullet}^2}{ak}-\frac{T^2}{N}$

④ $SSAB=\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{T_{ij\bullet}^2}{k}-\frac{T^2}{N}-(SSA+SSB)$

⑤ $SSW=SST-(SSA+SSB+SSW)=SST-\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{T_{ij\bullet}^2}{r}-\frac{T^2}{N}$

예제 383

4가지의 광고매체(인터넷, TV, 신문, 라디오) 와 3종류의 소비자층(10대, 20대, 30대)이 광고효과에 미치는 영향을 조사하고자 각 수준의 조합마다 3회씩 반복하여 얻은 자료가 다음과 같다. 분산분석표를 작성하고 각 요인과 교호작용에 대한 $\alpha=0.05$ 수준에서 검정을 하여라.

광고매체 / 소비자층	A_1	A_2	A_3	A_4	합계	평균
B_1	71 73 70	70 72 71	60 61 57	50 49 51	755	62.9167
B_2	79 75 76	75 74 76	68 67 65	59 60 55	829	69.0833
B_3	73 75 76	70 68 71	66 62 63	57 56 54	791	65.9167
합계	668	647	569	491	2375	
평균	74.2	71.887	63.222	54.556		65.972

(풀이)

$$\frac{T^2}{N}=\frac{(2375)^2}{36}=156684.0278,\ a=4,\ b=3,\ k=3$$

$$SST=\sum\sum\sum X_{ijk}^2-\frac{T^2}{N}=\{(71)^2+(73)^2+\ \cdots\ +(54)^2\}-156684.0278=2510.9722$$

$$SSA=\sum_{i=1}^{a}\frac{T_{i\bullet\bullet}^2}{bk}-\frac{T^2}{N}=\frac{1}{9}\{(668)^2+(647)^2+(569)^2+(491)^2\}-156684.0278=2168.75$$

$$SSB=\sum_{j=1}^{b}\frac{T_{\bullet j\bullet}^2}{ak}-\frac{T^2}{N}=\frac{1}{12}\{(755)^2+(829)^2+(791)^2\}-156684.0278=228.2222$$

$$SSAB=\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{T_{ij\bullet}^2}{k}-\frac{T^2}{N}-(SSA+SSB)$$

$$=\frac{1}{3}\{(214)^2+(230)^2+\ \cdots\ +(167)^2\}-156684.0278-(2168.75+228.2222)=44.6667$$

$$SSW=SST-(SSA+SSB+SSAB)$$

$$=2510.9722-(2168.75+228.2222+44.6667)=69.3333$$

<이원분산분석표>

분산원	제곱합	자유도	평균제곱	F값
인자 A	2168.75	3	722.9167	250.2394
인자 B	228.2222	2	114.1111	39.4998
교호작용	44.6667	6	7.4445	2.5769
오차	69.3333	24	2.8889	

① 독립변수 A효과에 대한 가설 검정

H_0 : 광고매체에 따라 차이가 없다. H_1 : 차이가 있다.

$F_{0.05}(3.24) = 3.01 < 250.2394$ 이므로 H_0를 기각한다.

∴ 광고매체에 따라 광고효과는 유의한 차이가 있다.

② 독립변수 B효과에 대한 가설 검정

H_0 : 소비자 층에 따라 차이가 없다. H_1 : 차이가 있다.

$F_{0.05}(2.24) = 3.40 < 39.4998$ 이므로 H_0를 기각한다.

∴ 소비자 층에 따라 광고효과는 유의한 차이가 있다.

③ 독립변수 A와 B의 상호작용효과에 대한 가설검정

H_0 : 두 변수 간에 상호작용이 없다. H_1 : 상호작용이 있다.

$F_{0.05}(6.24) = 2.51 < 2.5769$ 이므로 H_0를 기각한다.

∴ 광고매체와 소비자층 사이에 교호작용이 있다고 할 수 있다.