

이공학도를 위한 확률 및 통계학

-4장 수학적 기대값-

송 민 희(minhee@pel.sejong.ac.kr)

세종대학교 프로토콜공학연구실

목 차

- 보충
- 확률변수의 평균
- 분산과 공분산
- 선형결합된 확률변수의 평균과 분산
- 체비셰프 정리

목 차

- 보충
- 확률변수의 평균
- 분산과 공분산
- 선형결합된 확률변수의 평균과 분산
- 체비셰프 정리

보충

- 확률변수 $g(X)$ 의 분산 (1/2)

- 공식 (1/2)

- $\sigma^2 = E[g(X)^2] - \mu^2$

- 증명 (1/2)

* 확률변수 $g(X)$ 의 분산
 X 가 이산형인 경우:

$$\sigma_{g(X)}^2 = E\{[g(X) - \mu_{g(X)}]^2\} = \sum_x [g(x) - \mu_{g(X)}]^2 f(x)$$

X 가 연속형인 경우:

$$\sigma_{g(X)}^2 = E\{[g(X) - \mu_{g(X)}]^2\} = \int_{-\infty}^{\infty} [g(x) - \mu_{g(X)}]^2 f(x) dx$$

X 가 이산형인 경우

$$\begin{aligned}\sigma^2 &= \sum_x (g(x) - \mu)^2 f(x) = \sum_x (g(x)^2 - 2\mu g(x) + \mu^2) f(x) \\ &= \sum_x g(x)^2 f(x) - 2\mu \sum_x g(x) f(x) + \mu^2 \sum_x f(x)\end{aligned}$$

정의에 의해 $\mu = \sum_x g(x) f(x)$ 이고, $\sum_x f(x) = 1$

$$\therefore \sigma^2 = \sum_x g(x)^2 f(x) - 2\mu^2 + \mu^2 = \sum_x g(x)^2 f(x) - \mu^2 = E(g(X)^2) - \mu^2$$

보충

- 확률변수 $g(X)$ 의 분산 (2/2)

- 공식 (2/2)

- $\sigma^2 = E[g(X)^2] - \mu^2$

- 증명 (2/2)

* 확률변수 $g(X)$ 의 분산
 X 가 이산형인 경우:

$$\sigma_{g(X)}^2 = E\{[g(X) - \mu_{g(X)}]^2\} = \sum_x [g(x) - \mu_{g(X)}]^2 f(x)$$

X 가 연속형인 경우:

$$\sigma_{g(X)}^2 = E\{[g(X) - \mu_{g(X)}]^2\} = \int_{-\infty}^{\infty} [g(x) - \mu_{g(X)}]^2 f(x) dx$$

X 가 연속형인 경우

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (g(x) - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} (g(x)^2 - 2\mu x + \mu^2) f(x) dx \\ &= \int_{-\infty}^{\infty} g(x)^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} g(x) f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx\end{aligned}$$

정의에 의해 $\mu = \int_{-\infty}^{\infty} g(x) f(x) dx$ 이고, $\int_{-\infty}^{\infty} f(x) dx = 1$

$$\therefore \sigma^2 = \int_{-\infty}^{\infty} g(x)^2 f(x) dx - 2\mu^2 + \mu^2 = \int_{-\infty}^{\infty} g(x)^2 f(x) dx - \mu^2 = E(g(X)^2) - \mu^2$$

보충

- 상관계수(Correlation Coefficient) (1/4)

- 필요성

- 서로 다른 데이터쌍 간 선형적 관계의 강도 비교 가능
 - 정규화 된 값으로 정의되어 단위에 무관
- 두 변수 간 선형적 관계의 방향 파악 가능
 - 부호를 통해 두 변수 간 증가와 감소 방향 파악
- 두 변수가 얼마나 함께 변화하는지 정량적으로 평가 가능
 - 두 변수간 관계의 여부 뿐만 아니라 선형적 결합의 강도 표현

- 공식

- $$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

보충

• 상관계수 (2/4)

• 예제 4.16 (1/2)

	A	B	C	D
공부시간	1	2	3	4
시험점수	60	70	80	90

단위: 시간

단위: 점

	A	B	C	D
공부시간	60	120	180	240
시험점수	60	70	80	90

단위: 분

단위: 점

학생 A, B, C, D 의 공부 시간과 시험 점수 사이의 관계를 분석하려고 한다. 같은 학생들이 어떤 경우에는 공부시간을 시간(hour)으로, 다른 경우에는 분(minute)으로 기록했다. 두가지 경우에 대해 각각 공분산과 상관계수를 구하라.

- X: 공부시간, Y: 시험점수
- 시간 단위

- 평균: $\mu_X = \frac{1+2+3+4}{4} = 2.5$, $\mu_Y = \frac{60+70+80+90}{4} = 75$

- 공분산: $E(XY) = \frac{60+140+240+360}{4} = 200$, $\sigma_{XY} = 200 - (2.5)(75) = 12.5$

- 상관계수

$$E(X^2) = \frac{1^2+2^2+3^2+4^2}{4} = 7.5, \sigma_X^2 = 7.5 - (2.5)^2 = 1.25, \sigma_X = \sqrt{1.25} \approx 1.118$$

$$E(Y^2) = \frac{60^2+70^2+80^2+90^2}{4} = 5750, \sigma_Y^2 = 5750 - (75)^2 = 125, \sigma_Y = \sqrt{125} \approx 11.18$$

$$\rho_{XY} = \frac{12.5}{(1.118)(11.18)} = 1$$

보충

• 상관계수 (3/4)

• 예제 4.16 (2/2)

	A	B	C	D	
공부시간	1	2	3	4	단위: 시간
시험점수	60	70	80	90	단위: 점

	A	B	C	D	
공부시간	60	120	180	240	단위: 분
시험점수	60	70	80	90	단위: 점

학생 A, B, C, D 의 공부 시간과 시험 점수 사이의 관계를 분석하려고 한다. 같은 학생들이 어떤 경우에는 공부시간을 시간(hour)으로, 다른 경우에는 분(minute)으로 기록했다. 두가지 경우에 대해 각각 공분산과 상관계수를 구하라.

- X' : 공부시간, Y : 시험점수
- 분 단위

- 평균: $\mu_{X'} = \frac{60+120+180+240}{4} = 150$, $\mu_Y = 75$
- 공분산: $E(X'Y) = \frac{(60)(60)+(120)(70)+(180)(80)+(240)(90)}{4} = 12000$, $\sigma_{XY} = 12000 - (150)(75) = 750$
- 상관계수

$$E(X'^2) = \frac{60^2+120^2+180^2+240^2}{4} = 27000, \sigma_X^2 = 27000 - (150)^2 = 4500, \sigma_X = \sqrt{1.25} \approx 67.08$$

$$E(Y^2) = \frac{60^2+70^2+80^2+90^2}{4} = 5750, \sigma_Y^2 = 5750 - (75)^2 = 125, \sigma_Y = \sqrt{125} \approx 11.18$$

$$\rho_{XY} = \frac{750}{(67.08)(11.18)} = 1$$

보충

• 상관계수 (4/4)

• 예제 4.17

예제 4.14([표 3])에서 X 와 Y 의 상관계수를 구하라.

[표 3] X 와 Y 의 결합확률분포표

$f(x, y)$		x			$h(y)$
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
$g(x)$		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

- 예제 4.14에서 $\mu_X = \frac{3}{4}$, $\mu_Y = \frac{3}{14}$ 임을 확인

- X 와 Y 의 분산

$$E(X^2) = (0^2) \left(\frac{5}{14} \right) + (1^2) \left(\frac{15}{28} \right) + (2^2) \left(\frac{3}{28} \right) = \frac{27}{28}, \quad E(Y^2) = (0^2) \left(\frac{15}{28} \right) + (1^2) \left(\frac{3}{7} \right) + (2^2) \left(\frac{1}{28} \right) = \frac{4}{7}$$

$$\sigma_X^2 = \frac{27}{28} - \left(\frac{3}{4} \right)^2 = \frac{45}{112}, \quad \sigma_Y^2 = \frac{4}{7} - \left(\frac{1}{2} \right)^2 = \frac{9}{28}$$

- 예제 4.14에서 $\sigma_{XY} = -\frac{9}{56}$ 임을 확인

- X 와 Y 의 상관계수

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-\frac{9}{56}}{\sqrt{\left(\frac{45}{112} \right) \left(\frac{9}{28} \right)}} = -\frac{1}{\sqrt{5}} \approx -0.447$$

* 예제 4.14

어떤 상자에서 임의로 2개의 볼펜을 꺼낼 때, 청색볼펜의 수 X 와 적색볼펜의 수 Y 의 결합확률분포가 [표 3]과 같다. X 와 Y 의 공분산을 구하라.

보충

- 체비셰프 정리 (1/2)

- 필요성

- 실제 확률분포의 가정 없이 확률의 하한 제공
 - 평균과 분산만으로 확률변수가 특정 구간에 존재할 확률 하한 제시
- 평균 중심의 산포에 대한 확률적 보장 제공
 - 확률변수가 평균을 중심으로 일정 범위 내에 포함될 확률을 분산을 통해 정량적으로 보장
- 극단적 분포에 대해서도 성립하는 일반성 보유
 - 분포의 대칭성이나 정규성을 가정하지 않아도 모든 확률분포에 대해 항상 성립하는 보수적 기준을 제공

보충

• 체비셰프 정리 (2/2)

• 예제 4.27

어떤 전자기기의 배터리 수명 X (단위: 시간)에 대해 평균 수명이 10시간, 분산이 4일 때, 해당 배터리 수명이 6시간 이상 14시간 이하일 확률의 최소값을 구하라.

- $P(6 \leq X \leq 14)$
 $= P[10 - (2)(2) < X < 10 + (2)(2)] \geq 1 - \frac{1}{2^2} = \frac{3}{4}$
- $\therefore 75\%$

• 예제 4.28

어떤 웹서비스의 서버 응답 시간 X 를 측정하고 있다.

응답 시간의 정확한 분포 형태는 알 수 없으나, 장기간 로그 분석결과 평균 응답시간은 200ms, 분산은 10,000ms로 알려져 있다.

해당 웹서비스의 서버 응답 시간이 100ms 이상 300ms 이하일 확률의 최소값을 구하라.

- $\mu = 200ms, \quad \sigma^2 = 10000ms, \quad \sigma = 100ms$
- 확률의 범위

$$P(100 \leq X \leq 300) = P(|X - 200| < 100)$$

$$P(|X - 200| < (1)(100)) \geq 1 - \frac{1}{1^2} = 0$$

목 차

- 보충
- 확률변수의 평균
- 분산과 공분산
- 선형결합된 확률변수의 평균과 분산
- 체비셰프 정리

확률변수의 평균

- 산술 평균 (1/3)

- 정의

- 관측된 자료값들의 총합을 모든 자료값의 개수로 나눈 값

- 특징

- 이미 얻어진 표본값들을 이용하여 계산된 값
- 자료가 바뀌면 평균값도 함께 변화
- 전체 관측 횟수를 몰라도 값과 상대도수로 계산 가능

$$\begin{aligned}\text{평균} &= \frac{\sum (\text{값} \times \text{각 값의 관측횟수})}{\text{전체 관측횟수}} \\ &= \sum \text{값} \times \frac{\text{각 값의 관측횟수}}{\text{전체 관측횟수}} \\ &= \sum \text{값} \times \text{상대도수}\end{aligned}$$

확률변수의 평균

- 산술 평균 (2/3)

- 예제 4.1 (1/2)

한 쌍의 동전을 던져 앞면이 나오는 횟수에 대한 실험을 하고있다.

1) [표 1]은 16번의 실험을 통해 얻은 자료이다. 이에 대하여 앞면이 나오는 평균 횟수를 구하라.

[표 1] 예제 4.1의 자료

앞면의 개수	관측횟수
0	4
1	7
2	5

- 앞면에 나오는 평균 횟수

$$\begin{aligned} & \frac{(0)(4) + (1)(7) + (2)(5)}{16} \\ &= (0) \left(\frac{4}{16} \right) + (1) \left(\frac{7}{16} \right) + (2) \left(\frac{5}{16} \right) \\ &= 1.06 \end{aligned}$$

확률변수의 평균

- 산술 평균 (3/3)

- 예제 4.1 (2/2)

한 쌍의 동전을 던져 앞면이 나오는 횟수에 대한 실험을 하고있다.

2) 한 쌍의 동전을 던지는 실험에서 나타나는 표본공간을 정의하고, 1)의 계산식에서의 상대도수를 앞면이 나오는 개수별 확률로 대입하여 계산하라.

- 앞면: H , 뒷면: T
- 표본공간 $S = [HH, HT, TH, TT]$
- 각각의 경우가 나타날 확률: $\frac{1}{4}$
- 앞면이 나타날 개수 $X = \{0, 1, 2\}$ 에 대한 각 $P(X)$

$$P(X = 0) = P(TT) = \frac{1}{4}$$

$$P(X = 1) = P(TH) + P(HT) = \frac{1}{2}$$

$$P(X = 2) = P(HH) = \frac{1}{4}$$

- $(0) \left(\frac{1}{4}\right) + (1) \left(\frac{1}{2}\right) + (2) \left(\frac{1}{4}\right) = 1$

확률변수의 평균

- 수학적 기대값(Mathematical Expectation) (1/5)
 - 정의
 - 확률변수 X 가 취할 수 있는 값들에 대한 확률적 평균
 - 특징
 - 시행 횟수를 무한히 늘렸을 때 상대도수 평균이 수렴하는 값
 - 확률분포가 고정되면 기대값도 함께 고정
 - 관측된 표본값이 아닌 확률분포에 의해 정의됨
 - 필요성
 - 산술평균과 달리 각 결과에 대한 확률분포를 가중치로 반영
 - 현실 세계의 불확실한 결과 구조를 보다 잘 반영
 - 발생 확률은 낮으나 영향력이 큰 극단적인 결과까지 고려 가능

확률변수의 평균

- 수학적 기대값 (2/5)

- 표현

- 기대값(Expected Value)
- 확률변수 X 의 평균(Mean of the Random Variable X)
- X 의 확률분포의 평균(Mean of the Probability Distribution of X)
- $E(X)$
- μ_X 또는 μ
 - 어떤 확률 변수의 평균을 의미하는지가 명확할 때 X 생략

- 공식

- X 가 이산형인 경우: $\mu = E(X) = \sum_x xf(x)$
- X 가 연속형인 경우: $\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$

확률변수의 평균

- 수학적 기대값 (3/5)

- 예제 4.2

품질검사원이 7개의 부품으로 구성되어 있는 로트를 검사하려고한다. 만일 이 로트에 4개에 양호한 부품과 3개의 결함이 있는 부품이 들어 있다고 하면, 검사원이 3개의 부품을 추출하였을 때 나타나는 양호한 부품의 평균 개수를 구하라.

- X : 추출된 표본에 포함되는 양호한 부품의 수
- X 의 확률분포 $f(X)$

$$f(x) = \frac{\binom{4}{x} \binom{3}{3-x}}{\binom{7}{3}}, \quad x = 0, 1, 2, 3$$

- 각 개수가 나타날 확률

$$f(0) = \frac{1}{35}, f(1) = \frac{12}{35}, f(2) = \frac{18}{35}, f(3) = \frac{4}{35}$$

- 양호한 부품의 평균 개수

$$\mu = E(x) = (0) \left(\frac{1}{35} \right) + (1) \left(\frac{12}{35} \right) + (2) \left(\frac{18}{35} \right) + (3) \left(\frac{4}{35} \right) = \frac{12}{7} \doteq 1.7$$

확률변수의 평균

- 수학적 기대값 (4/5)

- 예제 4.3

의료기기 외판원이 어느 날 두 고객을 만나게 되었다. 첫 번째 고객과 거래가 성사될 가능성은 70%이고 이 경우 \$1000을 벌게 되며, 두 번째 고객과는 40%의 거래 성공 가능성에 성공 시 \$1500을 벌게 된다. 각 고객과의 거래 결과는 서로 독립적이라고 할 때 그가 기대할 수 있는 성공 보수는 얼마인가?

- 외판원이 얻을 수 있는 성공보수의 경우: \$0, \$1000, \$1500, \$2500
- 각 보수의 확률

$$f(\$0) = (1 - 0.7)(1 - 0.4) = 0.18$$

$$f(\$1000) = (0.7)(1 - 0.4) = 0.42$$

$$f(\$1500) = (1 - 0.7)(0.4) = 0.12$$

$$f(\$2500) = (0.7)(0.4) = 0.28$$

- 외판원의 평균 성공보수

$$E(X) = (\$0)(0.18) + (\$1000)(0.42) + (\$1500)(0.12) + (\$2500)(0.28) = \$1300$$

확률변수의 평균

- 수학적 기대값 (5/5)

- 예제 4.4

어떤 전자장치의 수명(단위: 시간)을 확률변수 X 라고 하자. 확률 밀도함수가

$$f(x) = \begin{cases} \frac{20,000}{x^3}, & x > 100 \\ 0, & \text{otherwise} \end{cases}$$

로 주어졌을 때 이 장치의 기대수명을 구하라.

- 전자장치의 기대 수명

$$\mu = E(X) = \int_{100}^{\infty} x \frac{20,000}{x^3} dx = \int_{100}^{\infty} \frac{20,000}{x^2} dx = 200$$

확률변수의 평균

- 확률변수 $g(X)$ 의 수학적 기대값 (1/3) * $g(X)$ 는 X 에 종속된 확률변수
- 정의
 - 확률변수 X 가 확률분포 $f(x)$ 를 가질 때 확률변수 $g(X)$ 의 기대값
- 특징
 - 원래의 값 X 가 아니라 X 로부터 계산되는 결과의 평균적 크기를 구하고자 할 때 사용
 - e.g., 비용, 손실, 제공, 효용 등
- 공식
 - X 가 이산형인 경우: $\mu_{g(X)} = E[g(X)] = \sum_x g(x)f(x)$
 - X 가 연속형인 경우: $\mu_{g(X)} = E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$

확률변수의 평균

- 확률변수 $g(X)$ 의 수학적 기대값 (2/3)

- 예제 4.5

어느 쾌청한 금요일 오후 4시에서 5시 사이에 세차장에서 서비스를 받는 차의 수를 X 라고 할 때 X 의 확률분포가 [표 2]와 같다고 하자.

$g(X) = 2X - 1$ 을 종업원이 받는 수당(단위: 달러)이라고 할 때, 이 시간대의 종업원의 기대수익을 구하라.

[표 2] 예제 4.5의 자료

x	4	5	6	7	8	9
$P(X = x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$

- 종업원의 기대수익

$$\begin{aligned} E[g(X)] &= E(2X - 1) = \sum_{x=4}^9 (2x - 1)f(x) \\ &= (7) \left(\frac{1}{12} \right) + (9) \left(\frac{1}{12} \right) + (11) \left(\frac{1}{4} \right) + (13) \left(\frac{1}{4} \right) + (15) \left(\frac{1}{6} \right) + (17) \left(\frac{1}{6} \right) \\ &= \$12.67 \end{aligned}$$

확률변수의 평균

- 확률변수 $g(X)$ 의 수학적 기대값 (3/3)
- 예제 4.6

확률변수 X 의 밀도함수가

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

일 때 $g(X) = 4X + 3$ 의 기대값을 구하라.

- $g(X)$ 의 기대값

$$E[g(X)] = E(4X + 3) = \int_{-1}^2 \frac{(4x + 3)x^2}{3} dx = \frac{1}{3} \int_{-1}^2 (4x^3 + 3x^2) dx = 8$$

확률변수의 평균

- 확률변수 $g(X, Y)$ 의 수학적 기대값 (1/3)

- 정의

- 확률변수 X, Y 가 결합확률분포 $f(x, y)$ 를 가질 때 확률변수 $g(X, Y)$ 의 기대값

- 특징

- 두 확률변수 X, Y 가 결합하여 함께 계산되는 결과의 평균적 크기를 구하고자 할 때 사용
 - e.g., 총비용, 총 손실, 거리, 상호작용 효과 등

- 공식

- X 와 Y 가 이산형인 경우:

$$\mu_{g(X,Y)} = E[g(X,Y)] = \sum_x \sum_y g(x, y) f(x, y)$$

- X 와 Y 가 연속형인 경우:

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

확률변수의 평균

- 확률변수 $g(X, Y)$ 의 수학적 기대값 (2/3)

- 예제 4.7

X 와 Y 를 [표 3]과 같은 결합확률분포를 가지는 확률변수라 할 때, $g(X, Y) = XY$ 의 기대값을 구하라.

- $g(X, Y)$ 의 기대값

$$\begin{aligned} E(XY) &= \sum_{x=0}^2 \sum_{y=0}^2 xyf(x, y) \\ &= (0)(0)f(0,0) + (0)(1)f(0,1) + (1)(0)f(1,0) + (1)(1)f(1,1) + (2)(0)f(2,0) \\ &= f(1,1) = \frac{3}{14} \end{aligned}$$

[표 3] X 와 Y 의 결합확률분포표

$f(x, y)$		x			행의 합
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
열의 합		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

확률변수의 평균

- 확률변수 $g(X, Y)$ 의 수학적 기대값 (3/3)
- 예제 4.8

다음과 같은 결합밀도함수에 대하여 $E\left(\frac{Y}{X}\right)$ 를 구하라.

$$f(x, y) = \begin{cases} \frac{x(1 + 3y^2)}{4}, & 0 < x < 2, 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

- $g(X, Y) = \frac{Y}{X}$
- $E\left(\frac{Y}{X}\right)$

$$E\left(\frac{Y}{X}\right) = \int_0^1 \int_0^2 \frac{y(1 + 3y^2)}{4} dx dy = \int_0^1 \frac{y + 3y^3}{2} dy = \frac{5}{8}$$

목 차

- 보충
- 확률변수의 평균
- 분산과 공분산
- 선형결합된 확률변수의 평균과 분산
- 체비셰프 정리

분산과 공분산

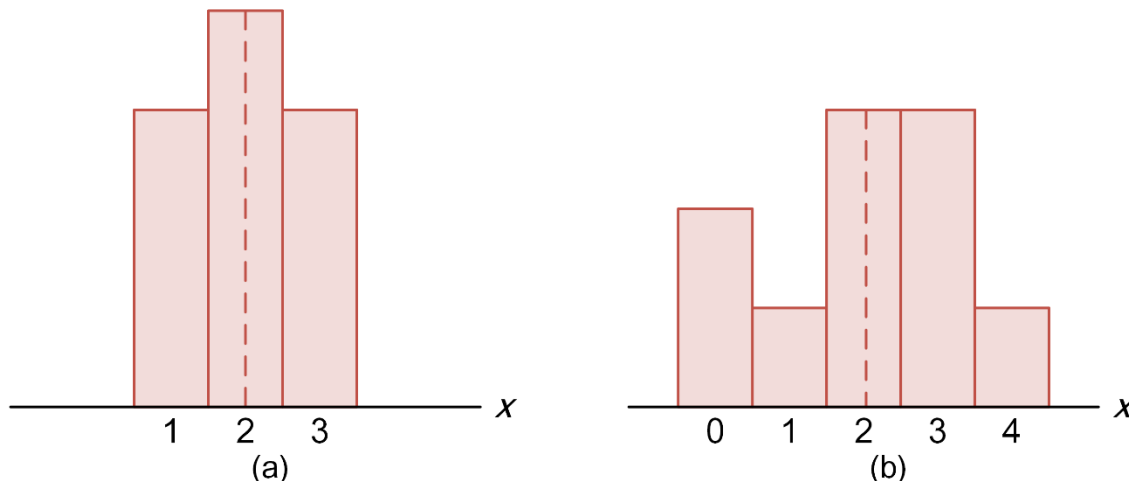
- 분포의 산포(Variability)

- 정의

- 확률변수가 중심위치로부터 얼마나 흩어져 있는 지를 나타내는 정도

- 필요성

- 수학적 기대값만으로는 확률변수의 분포 형태와 변동 정도를 충분히 설명할 수 없음
 - e.g., 평균은 $\mu = 2$ 로 같지만, 산포가 다른 두 분포



<그림 1> 평균은 같고 산포는 다른 분포 그래프

분산과 공분산

- 확률변수 X 의 분산(Variance of the Random Variable X) (1/6)
 - 정의
 - 확률변수 X 가 그 평균(μ_X)으로부터 얼마나 벗어나 있는지를 나타내는 정도의 평균
 - 특징
 - x 값들이 μ 에 가까이 놓여있는 경우, μ 로 부터 떨어져 있는 경우보다 작은 값을 가짐
 - 표현
 - X 의 확률분포의 분산(Variance of the Probability Distribution of X)
 - σ_X^2 또는 σ^2
 - 어떤 확률 변수의 평균을 의미하는지가 명확할 때 X 생략

분산과 공분산

- 확률변수 X 의 분산 (2/6)

- 공식 (1/2)

- X 가 이산형일 때:

$$\sigma^2 = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x)$$

- X 가 연속형일 때:

$$\sigma^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

* $x - \mu$: 관측값의 평균으로부터의 편차(Deviation)

* σ : X 의 표준편차(Standard Deviation), 분산의 양의 제곱근

분산과 공분산

- 확률변수 X 의 분산 (3/6)

- 공식 (2/2)

- $\sigma^2 = E(X^2) - \mu^2$

- 증명

* $x - \mu$: 관측값의 평균으로부터의 편차(Deviation)

* σ : X 의 표준편차(Standard Deviation), 분산의 양의 제곱근

X 가 이산형인 경우

$$\sigma^2 = \sum_x (x - \mu)^2 f(x) = \sum_x (x^2 - 2\mu x + \mu^2) f(x) = \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x)$$

정의에 의해 $\mu = \sum_x x f(x)$ 이고, $\sum_x f(x) = 1$

$$\therefore \sigma^2 = \sum_x x^2 f(x) - 2\mu^2 + \mu^2 = \sum_x x^2 f(x) - \mu^2 = E(X^2) - \mu^2$$

X 가 연속형인 경우

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} (x^2 - 2\mu x + \mu^2) f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} x f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx$$

정의에 의해 $\mu = \int_{-\infty}^{\infty} x f(x) dx$ 이고, $\int_{-\infty}^{\infty} f(x) dx = 1$

$$\therefore \sigma^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - 2\mu^2 + \mu^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = E(X^2) - \mu^2$$

분산과 공분산

• 확률변수 X 의 분산 (4/6)

• 예제 4.9

[표 4] 예제 4.9의 A 회사의 확률분포

x	1	2	3
$f(x)$	0.3	0.4	0.3

[표 5] 예제 4.9의 B 회사의 확률분포

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

A와 B 두 회사에서 어느 날 사업목적으로 사용된 자동차 수를 확률변수 X 라고 하자. A 회사에 대한 확률분포와 B 회사에 대한 확률분포가 각각 [표 4], [표 5]와 같을 때, B 회사에 대한 확률분포의 분산이 A 회사의 분산보다 큼을 보여라.

• A 회사

$$\mu_A = E(X) = (1)(0.3) + (2)(0.4) + (3)(0.3) = 2.0$$

$$\sigma_A^2 = \sum_{x=1}^3 (x - 2)^2 f(x) = (1 - 2)^2(0.3) + (2 - 2)^2(0.4) + (3 - 2)^2(0.3) = 0.6$$

• B 회사

$$\mu_B = E(X) = (0)(0.2) + (1)(0.1) + (2)(0.3) + (3)(0.3) + (4)(0.1) = 2.0$$

$$\begin{aligned}\sigma_B^2 &= \sum_{x=0}^4 (x - 2)^2 f(x) \\ &= (0 - 2)^2(0.2) + (1 - 2)^2(0.1) + (2 - 2)^2(0.3) + (3 - 2)^2(0.3) + (4 - 2)^2(0.1) \\ &= 1.6\end{aligned}$$

분산과 공분산

- 확률변수 X 의 분산 (5/6)

- 예제 4.10

[표 6] 예제 4.10의 자료

x	0	1	2	3
$f(x)$	0.51	0.38	0.10	0.01

생산라인으로부터 3개의 부품을 추출하여 검사하였을 때 결함이 있는 부품의 수를 확률변수 X 라고 하자. X 의 확률분포가 [표 6]과 같을 때 확률변수 X 의 분산 공식 ' $\sigma^2 = E(X^2) - \mu^2$ '을 이용하여 σ^2 을 계산하라.

- X 의 평균

$$\mu = (0)(0.51) + (1)(0.38) + (2)(0.10) + (3)(0.01) = 0.61$$

- $E(X^2)$

$$E(X^2) = (0)(0.51) + (1)(0.38) + (4)(0.10) + (9)(0.01) = 0.87$$

- X 의 분산

$$\sigma^2 = 0.87 - (0.61)^2 = 0.4979$$

분산과 공분산

- 확률변수 X 의 분산 (6/6)

- 예제 4.11

어느 연쇄점의 주당 콜라의 수요(단위: 1000리터)가 다음과 같은 확률 분포를 가지는 확률변수 X 라고 할 때, X 의 평균과 분산을 구하라.

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

- X 의 평균

$$\mu = E(X) = 2 \int_1^2 x(x-1) dx = \frac{5}{3}$$

- $E(X^2)$

$$E(X^2) = 2 \int_1^2 x^2(x-1) dx = \frac{17}{6}$$

- X 의 분산

$$\sigma^2 = \frac{17}{6} - \left(\frac{5}{3}\right)^2 = \frac{1}{18}$$

분산과 공분산

- 확률변수 $g(X)$ 의 분산 (1/5) * $g(X)$ 는 X 에 종속된 확률변수
- 정의
 - 확률변수 X 가 확률분포 $f(x)$ 를 가질 때 $g(X)$ 가 그 기대값($\mu_{g(X)}$)으로부터 벗어난 정도의 제곱의 평균

- 공식 (1/3)

- X 가 이산형인 경우:

$$\sigma_{g(X)}^2 = E \left\{ [g(X) - \mu_{g(X)}]^2 \right\} = \sum_x [g(x) - \mu_{g(X)}]^2 f(x)$$

- X 가 연속형인 경우:

$$\sigma_{g(X)}^2 = E \left\{ [g(X) - \mu_{g(X)}]^2 \right\} = \int_{-\infty}^{\infty} [g(x) - \mu_{g(X)}]^2 f(x) dx$$

분산과 공분산

- 확률변수 $g(X)$ 의 분산 (2/5)

- 공식 (2/3)

- $\sigma^2 = E[g(X)^2] - \mu^2$

- 증명 (2/2)

X 가 이산형인 경우

$$\begin{aligned}\sigma^2 &= \sum_x (g(x) - \mu)^2 f(x) = \sum_x (g(x)^2 - 2\mu g(x) + \mu^2) f(x) \\ &= \sum_x g(x)^2 f(x) - 2\mu \sum_x g(x) f(x) + \mu^2 \sum_x f(x)\end{aligned}$$

정의에 의해 $\mu = \sum_x g(x) f(x)$ 이고, $\sum_x f(x) = 1$

$$\therefore \sigma^2 = \sum_x g(x)^2 f(x) - 2\mu^2 + \mu^2 = \sum_x g(x)^2 f(x) - \mu^2 = E(g(X)^2) - \mu^2$$

분산과 공분산

- 확률변수 $g(X)$ 의 분산 (3/5)

- 공식 (3/3)

- $\sigma^2 = E[g(X)^2] - \mu^2$

- 증명 (2/2)

X 가 연속형인 경우

$$\begin{aligned}\sigma^2 &= \int_{-\infty}^{\infty} (g(x) - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} (g(x)^2 - 2\mu x + \mu^2) f(x) dx \\ &= \int_{-\infty}^{\infty} g(x)^2 f(x) dx - 2\mu \int_{-\infty}^{\infty} g(x) f(x) dx + \mu^2 \int_{-\infty}^{\infty} f(x) dx\end{aligned}$$

정의에 의해 $\mu = \int_{-\infty}^{\infty} g(x) f(x) dx$ 이고, $\int_{-\infty}^{\infty} f(x) dx = 1$

$$\therefore \sigma^2 = \int_{-\infty}^{\infty} g(x)^2 f(x) dx - 2\mu^2 + \mu^2 = \int_{-\infty}^{\infty} g(x)^2 f(x) dx - \mu^2 = E(g(X)^2) - \mu^2$$

분산과 공분산

- 확률변수 $g(X)$ 의 분산 (4/5)

[표 7] 예제 4.12의 자료

x	0	1	2	3
$f(x)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{8}$

- 예제 4.12

확률변수 X 의 확률분포가 [표 7]과 같을 때 $g(X) = 2X + 3$ 의 분산을 계산하라.

- $g(X)$ 의 평균

$$\mu_{2X+3} = E(2X + 3) = \sum_{x=0}^3 (2x + 3)f(x) = 6$$

- $g(X)$ 의 분산

$$\begin{aligned}\sigma_{2X+3}^2 &= E\{[(2X + 3) - \mu_{2X+3}]^2\} = E\{(2X + 3 - 6)^2\} \\ &= E(4X^2 - 12X + 9) = \sum_{x=0}^3 (4x^2 - 12x + 9)f(x) = 4\end{aligned}$$

분산과 공분산

- 확률변수 $g(X)$ 의 분산 (5/5)
- 예제 4.13

X 를 예제 4.6의 확률밀도함수를 가지는 확률변수라고 할 때, 확률변수 $g(X) = 4X + 3$ 의 분산을 구하라.

- 예제 4.6에서 $\mu_{4X+3} = 8$ 임을 확인
- $g(X)$ 의 분산

$$\begin{aligned}\sigma_{4X+3}^2 &= E\{[(4X + 3) - 8]^2\} = E[(4X - 5)^2] \\ &= \int_{-1}^2 (4x - 5)^2 \frac{x^2}{3} dx = \frac{1}{3} \int_{-1}^2 (16x^4 - 40x^3 + 25x^2) dx = \frac{51}{5}\end{aligned}$$

*예제 4.6의 확률밀도함수 $f(x)$

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

분산과 공분산

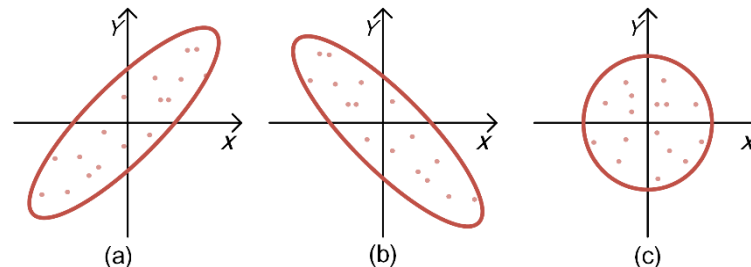
- 공분산(Covariance) (1/6)

- 정의

- 평균을 기준으로 두 확률변수의 선형적 변동관계를 나타내는 척도

- 특징

- 두 확률변수 사이의 선형적 관계만을 측정
- 두 확률변수가 함께 증가하거나 감소하는 경향을 측정
 - 양의 값: 같은 방향으로 변동 - <그림 2> (a)
 - 음의 값: 반대 방향으로 변동 - <그림 2> (b)
 - 0: 선형적 관계가 없음 - <그림 2> (c)
 - 비선형적 관계나 독립성 여부를 판단할 수 없음



<그림 2> 공분산의 부호에 따른 두 확률변수의 선형적 관계

분산과 공분산

- 공분산 (2/6)

- 공식 (1/3)

- X 와 Y 가 이산형인 경우:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x, y)$$

- X 와 Y 가 연속형인 경우:

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y) dx dy$$

분산과 공분산

- 공분산 (3/6)

- 공식 (2/3)

- μ_X 와 μ_Y 를 평균으로 하는 두 확률변수 X 와 Y 의 공분산인 경우:

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y$$

- 증명 (1/2)

이산형인 경우

$$\begin{aligned}\sigma_{XY} &= \sum_x \sum_y (x - \mu_X)(y - \mu_Y)f(x, y) \\ &= \sum_x \sum_y xyf(x, y) - \mu_X \sum_x \sum_y yf(x, y) - \mu_Y \sum_x \sum_y xf(x, y) + \mu_X\mu_Y \sum_x \sum_y f(x, y)\end{aligned}$$

정의에 의해서 $\mu_X = \sum_x xf(x, y)$, $\mu_Y = \sum_y yf(x, y)$ 이고, $\sum_x \sum_y f(x, y) = 1$

$$\therefore \sigma_{XY} = E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y$$

분산과 공분산

- 공분산 (4/6)

- 공식 (3/3)

- μ_X 와 μ_Y 를 평균으로 하는 두 확률변수 X 와 Y 의 공분산인 경우:

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y$$

- 증명 (2/2)

연속형인 경우

$$\begin{aligned}\sigma_{XY} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x, y) dx dy \\ &= \iint xyf(x, y) dx dy - \mu_X \iint yf(x, y) dx dy - \mu_Y \iint xf(x, y) dx dy + \mu_X\mu_Y \iint f(x, y) dx dy\end{aligned}$$

정의에 의해서 $\mu_X = \iint xf(x, y) dx dy$, $\mu_Y = \iint yf(x, y) dx dy$ 이고, $\iint f(x, y) dx dy = 1$

$$\therefore \sigma_{XY} = E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y$$

분산과 공분산

• 공분산 (5/6)

• 예제 4.14

어떤 상자에서 임의로 2개의 볼펜을 꺼낼 때, 청색볼펜의 수 X 와 적색볼펜의 수 Y 의 결합확률분포가 예제 4.7의 [표 3]과 같다. X 와 Y 의 공분산을 구하라.

[표 3] X 와 Y 의 결합확률분포표

$f(x, y)$		x			$h(y)$
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
$g(x)$		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

- 예제 4.7에서 $E(XY) = \frac{3}{14}$ 임을 확인
- X 와 Y 에 대한 평균

$$\mu_X = \sum_{x=0}^2 xg(x) = (0) \left(\frac{5}{14} \right) + (1) \left(\frac{15}{28} \right) + (2) \left(\frac{3}{28} \right) = \frac{3}{4}$$

$$\mu_Y = \sum_{y=0}^2 yh(y) = (0) \left(\frac{15}{28} \right) + (1) \left(\frac{3}{7} \right) + (2) \left(\frac{1}{28} \right) = \frac{1}{2}$$

- X 와 Y 의 공분산

$$\sigma_{XY} = E(XY) - \mu_X \mu_Y = \frac{3}{14} - \left(\frac{3}{4} \right) \left(\frac{1}{2} \right) = -\frac{9}{56}$$

분산과 공분산

• 공분산 (6/6)

• 예제 4.15

마라톤코스를 완주한 남자의 비율 X , 여자의 비율 Y 의 결합확률분포가 다음과 같다. X 와 Y 의 공분산을 구하라.

$$f(x, y) = \begin{cases} 8xy, & 0 \leq y \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

- 남자의 비율에 대한 주변밀도함수 $g(x)$, 여자의 비율에 대한 주변밀도함수 $h(y)$

$$g(x) = \begin{cases} 4x^3, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}, \quad h(y) = \begin{cases} 4y(1 - y^2), & 0 \leq y \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

- X 와 Y 의 평균

$$\mu_X = E(X) = \int_0^1 xg(x) dx = \int_0^1 4x^4 dx = \frac{4}{5} \quad \mu_Y = \int_0^1 yh(y) dy = \int_0^1 4y^2(1 - y^2) dy = \frac{8}{15}$$

- X 와 Y 의 공분산

$$E(XY) = \int_0^1 \int_y^1 8x^2y^2 dx dy = \frac{4}{9}$$

$$\sigma_{XY} = E(XY) - \mu_X\mu_Y = \frac{4}{9} - \left(\frac{4}{5}\right)\left(\frac{8}{15}\right) = \frac{4}{225}$$

분산과 공분산

- 상관계수(Correlation Coefficient) (1/6)

- 정의

- 평균을 기준으로 두 확률변수 사이의 선형적 변동 관계의 강도와 방향을 나타내는 척도

- 특징

- 공분산을 표준편차로 정규화한 값
 - 공분산이 0이면 상관계수값도 0 (단, 표준편차 $\neq 0$)
- 확률변수의 단위에 무관
- 값의 범위가 항상 $-1 \leq \rho_{XY} \leq 1$

분산과 공분산

- 상관계수 (2/6)

- 필요성

- 서로 다른 데이터셋 간 선형적 관계의 강도 비교 가능
 - 정규화 된 값으로 정의되어 단위에 무관
- 두 변수 간 선형적 관계의 방향 파악 가능
 - 부호를 통해 두 변수 간 증가와 감소 방향 파악
- 두 변수가 얼마나 함께 변화하는지 정량적으로 평가 가능
 - 두 변수간 관계의 여부 뿐만 아니라 선형적 결합의 강도 표현

- 공식

- $$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

분산과 공분산

• 상관계수 (3/6)

• 예제 4.16 (1/2)

	A	B	C	D
공부시간	1	2	3	4
시험점수	60	70	80	90

단위: 시간

단위: 점

	A	B	C	D
공부시간	60	120	180	240
시험점수	60	70	80	90

단위: 분

단위: 점

학생 A, B, C, D 의 공부 시간과 시험 점수 사이의 관계를 분석하려고 한다. 같은 학생들이 어떤 경우에는 공부시간을 시간(hour)으로, 다른 경우에는 분(minute)으로 기록했다. 두가지 경우에 대해 각각 공분산과 상관계수를 구하라.

- X: 공부시간, Y: 시험점수
- 시간 단위

- 평균: $\mu_X = \frac{1+2+3+4}{4} = 2.5$, $\mu_Y = \frac{60+70+80+90}{4} = 75$

- 공분산: $E(XY) = \frac{60+140+240+360}{4} = 200$, $\sigma_{XY} = 200 - (2.5)(75) = 12.5$

- 상관계수

$$E(X^2) = \frac{1^2+2^2+3^2+4^2}{4} = 7.5, \sigma_X^2 = 7.5 - (2.5)^2 = 1.25, \sigma_X = \sqrt{1.25} \approx 1.118$$

$$E(Y^2) = \frac{60^2+70^2+80^2+90^2}{4} = 5750, \sigma_Y^2 = 5750 - (75)^2 = 125, \sigma_Y = \sqrt{125} \approx 11.18$$

$$\rho_{XY} = \frac{12.5}{(1.118)(11.18)} = 1$$

분산과 공분산

• 상관계수 (4/6)

• 예제 4.16 (2/2)

	A	B	C	D
공부시간	1	2	3	4
시험점수	60	70	80	90

단위: 시간

단위: 점

	A	B	C	D
공부시간	60	120	180	240
시험점수	60	70	80	90

단위: 분

단위: 점

학생 A, B, C, D 의 공부 시간과 시험 점수 사이의 관계를 분석하려고 한다. 같은 학생들이 어떤 경우에는 공부시간을 시간(hour)으로, 다른 경우에는 분(minute)으로 기록했다. 두가지 경우에 대해 각각 공분산과 상관계수를 구하라.

• X' : 공부시간, Y : 시험점수

• 분 단위

• 평균: $\mu_{X'} = \frac{60+120+180+240}{4} = 150$, $\mu_Y = 75$

• 공분산: $E(X'Y) = \frac{(60)(60)+(120)(70)+(180)(80)+(240)(90)}{4} = 12000$, $\sigma_{XY} = 12000 - (150)(75) = 750$

• 상관계수

$$E(X'^2) = \frac{60^2+120^2+180^2+240^2}{4} = 27000, \sigma_X^2 = 27000 - (150)^2 = 4500, \sigma_X = \sqrt{1.25} \approx 67.08$$

$$E(Y^2) = \frac{60^2+70^2+80^2+90^2}{4} = 5750, \sigma_Y^2 = 5750 - (75)^2 = 125, \sigma_Y = \sqrt{125} \approx 11.18$$

$$\rho_{XY} = \frac{750}{(67.08)(11.18)} = 1$$

분산과 공분산

• 상관계수 (5/6)

• 예제 4.17

예제 4.14([표 3])에서 X 와 Y 의 상관계수를 구하라.

[표 3] X 와 Y 의 결합확률분포표

$f(x, y)$		x			$h(y)$
		0	1	2	
y	0	$\frac{3}{28}$	$\frac{9}{28}$	$\frac{3}{28}$	$\frac{15}{28}$
	1	$\frac{3}{14}$	$\frac{3}{14}$	0	$\frac{3}{7}$
	2	$\frac{1}{28}$	0	0	$\frac{1}{28}$
$g(x)$		$\frac{5}{14}$	$\frac{15}{28}$	$\frac{3}{28}$	1

- 예제 4.14에서 $\mu_X = \frac{3}{4}$, $\mu_Y = \frac{3}{14}$ 임을 확인

- X 와 Y 의 분산

$$E(X^2) = (0^2) \left(\frac{5}{14} \right) + (1^2) \left(\frac{15}{28} \right) + (2^2) \left(\frac{3}{28} \right) = \frac{27}{28}, \quad E(Y^2) = (0^2) \left(\frac{15}{28} \right) + (1^2) \left(\frac{3}{7} \right) + (2^2) \left(\frac{1}{28} \right) = \frac{4}{7}$$

$$\sigma_X^2 = \frac{27}{28} - \left(\frac{3}{4} \right)^2 = \frac{45}{112}, \quad \sigma_Y^2 = \frac{4}{7} - \left(\frac{1}{2} \right)^2 = \frac{9}{28}$$

- 예제 4.14에서 $\sigma_{XY} = -\frac{9}{56}$ 임을 확인

- X 와 Y 의 상관계수

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{-\frac{9}{56}}{\sqrt{\left(\frac{45}{112} \right) \left(\frac{9}{28} \right)}} = -\frac{1}{\sqrt{5}} \approx -0.447$$

분산과 공분산

- 상관계수 (6/6)

- 예제 4.18

예제 4.15에서 X 와 Y 의 상관계수를 구하라.

*예제 4.15의 확률밀도함수 $f(x, y)$

$$f(x, y) = \begin{cases} 8xy, & 0 \leq y \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

- 예제 4.15에서 $\mu_X = \frac{4}{5}$, $\mu_Y = \frac{8}{15}$ 임을 확인
- X 와 Y 의 분산

$$E(X^2) = \int_0^1 4x^5 dx = \frac{2}{3}, \quad E(Y^2) = \int_0^1 4y^3(1-y^2)dx = 1 - \frac{2}{3} = \frac{1}{3}$$

$$\sigma_X^2 = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75}, \quad \sigma_Y^2 = \frac{1}{3} - \left(\frac{8}{15}\right)^2 = \frac{11}{225}$$

- 예제 4.15에서 $\sigma_{XY} = \frac{4}{225}$ 임을 확인
- X 와 Y 의 상관계수

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{4}{225}}{\sqrt{\left(\frac{2}{75}\right)\left(\frac{11}{225}\right)}} = \frac{4}{\sqrt{66}}$$

목 차

- 보충
- 확률변수의 평균
- 분산과 공분산
- 선형결합된 확률변수의 평균과 분산
- 체비셰프 정리

선형결합된 확률변수의 평균과 분산

- 선형결합

- 정의

- 여러 변수에 상수를 곱한 후 그 합으로 나타낸 형태

- 표현

- $a_1v_1 + a_2v_2 + \dots + a_nv_n$ ($a_1, a_2, \dots, a_n$ 은 상수)

- 선형결합된 확률변수

- 정의

- 여러 확률변수에 상수를 곱한 후 더하여 정의된 새로운 확률변수

- 표현

- 확률변수 $X_1, X_2, \dots, X_n$ 과 상수 $a_1, a_2, \dots, a_n$ 에 대해
$$Y = a_1X_1 + a_2X_2 + \dots + a_nX_n$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (1/9)
- a 와 b 가 상수이면 $E(aX + b) = aE(X) + b$

- 증명

기대값의 정의에 따라

$$E(aX + b) = \int_{-\infty}^{\infty} (ax + b)f(x) dx = a \int_{-\infty}^{\infty} xf(x) dx + b \int_{-\infty}^{\infty} f(x) dx$$

$\int_{-\infty}^{\infty} xf(x) dx = E(X)$ 이고, $\int_{-\infty}^{\infty} f(x) dx = 1$ 이므로 $E(aX + b) = aE(X) + b$

- 따름정리 4.1

$a = 0$ 으로 놓으면 $E(b) = b$ 가 된다.

- 따름정리 4.2

$b = 0$ 으로 놓으면 $E(aX) = aE(X)$ 가 된다.

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (2/9)

- 예제 4.19

' a 와 b 가 상수이면 $E(aX + b) = aE(X) + b$ '임을 이산형 확률변수 $g(X) = 2X - 1$ 에 적용하여 예제 4.5를 다시 풀어 보라.

▶ 예제 4.5

어느 쾌청한 금요일 오후 4시에서 5시 사이에 세차장에서 서비스를 받는 차의 수를 X 라고 할 때 X 의 확률분포가 [표 2]와 같다고 하자. $g(X) = 2X - 1$ 이 종업원이 받는 수당(단위: 달러)일 때, 이 시간대의 종업원의 기대수익을 구하라.

[표 2] 예제 4.5의 자료

x	4	5	6	7	8	9
$P(X = x)$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{6}$

- $E[g(X)]$

$$E(2X - 1) = 2E(X) - 1$$

- $E(X)$

$$\mu = E(X) = \sum_{x=4}^9 xf(x) = (4)\left(\frac{1}{12}\right) + (5)\left(\frac{1}{12}\right) + (6)\left(\frac{1}{4}\right) + (7)\left(\frac{1}{4}\right) + (8)\left(\frac{1}{6}\right) + (9)\left(\frac{1}{6}\right) = \frac{41}{6}$$

- 종업원의 기대수익

$$\mu_{2X-1} = (2)\left(\frac{41}{6}\right) - 1 \doteq \$12.6$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (3/9)

- 예제 4.20

' a 와 b 가 상수이면 $E(aX + b) = aE(X) + b$ '임을 연속형 확률변수 $g(X) = 4X + 3$ 에 적용하여 예제 4.6를 다시 풀어 보라.

▶ 예제 4.6

확률변수 X 의 밀도함수가 아래와 같을 때 $g(X) = 4X + 3$ 의 기대값을 구하라.

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

- $E[g(x)]$

$$E(4X + 3) = 4E(X) + 3$$

- $E(X)$

$$\mu = E(X) = \int_{-1}^2 x \left(\frac{x^2}{3} \right) dx = \int_{-1}^2 \frac{x^3}{3} dx = \frac{5}{4}$$

- 최종 기대값

$$\mu_{4X+3} = E(4X + 3) = (4) \left(\frac{5}{4} \right) + 3 = 8$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (4/9)
- 두 개이상의 확률변수 X 의 함수의 합이나 차의 기대값은 각 함수의 기대값의 합이나 차와 동일
: $E[g(X) \pm h(X)] = E[g(X)] \pm E[h(X)]$

- 증명

정의에 의하여

$$\begin{aligned} E[g(X) \pm h(X)] &= \int_{-\infty}^{\infty} [g(x) \pm h(x)]f(x) dx \\ &= \int_{-\infty}^{\infty} g(x)f(x) dx \pm \int_{-\infty}^{\infty} h(x)f(x) dx \\ &= E[g(X)] \pm E[h(X)] \end{aligned}$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (5/9)
- 예제 4.21

X 가 [표 8]과 같은 확률분포를 가지는 확률변수일 때, $Y = (X - 1)^2$ 의 기대값을 구하라.

- $E(Y)$

$$E[(X - 1)^2] = E(X^2 - 2X + 1) = E(X^2) - 2E(X) + E(1)$$

- 따름정리 4.1에 의해 $E(1) = 1$
- $E(X)$ 와 $E(X^2)$

$$E(X) = (0)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{2}\right) + (2)(0) + (3)\left(\frac{1}{6}\right) = 1$$

$$E(X^2) = (0)\left(\frac{1}{3}\right) + (1)\left(\frac{1}{2}\right) + (4)(0) + (9)\left(\frac{1}{6}\right) = 2$$

- 최종 기대값

$$E[(X - 1)^2] = 2 - (2)(1) + 1 = 1$$

[표 8] 예제 4.20의 자료

x	0	1	2	3
$f(x)$	$\frac{1}{3}$	$\frac{1}{2}$	0	$\frac{1}{6}$

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (6/9)

- 예제 4.22

어느 연쇄점에서 판매하는 어떤 음료의 주당 수요(단위: 1000리터)가 연속형 확률변수 $g(X) = X^2 + X - 2$ 로 나타나며, X 는 다음과 같은 확률분포를 가진다고 할 때, 음료수의 주당 수요의 기대값을 구하라.

$$f(x) = \begin{cases} 2(x-1), & 1 < x < 2 \\ 0, & \text{otherwise} \end{cases}$$

- $E[g(X)]$

$$E(X^2 + X - 2) = E(X^2) + E(X) - E(2)$$

- 따름정리 4.1에 의해 $E(2) = 2$

- $E(X)$ 와 $E(X^2)$

$$E(X) = \int_1^2 2x(x-1) dx = 2 \int_1^2 (x^2 - x) dx = \frac{5}{3}$$

$$E(X^2) = \int_1^2 2x^2(x-1) dx = 2 \int_1^2 (x^3 - x^2) dx = \frac{17}{6}$$

- 주당 평균 수요

$$E(X^2 + X - 2) = \frac{17}{6} + \frac{5}{3} - 2 = \frac{5}{2}, \quad 1000 \times \frac{5}{2} = 2500$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (7/9)

- 두 확률변수 X 와 Y 의 함수들의 합이나 차의 기대값은 각 함수의 기대값의 합이나 차와 동일

$$: E[g(X, Y) \pm h(X, Y)] = E[g(X, Y)] \pm E[h(X, Y)]$$

- 증명

확률변수 $g(x, y)$ 의 수학적 기대값에 대한 정의에 의해서

$$\begin{aligned} E[g(X, Y) \pm h(X, Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [g(x, y) \pm h(x, y)] f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy \pm \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y) dx dy \\ &= E[g(X, Y)] \pm E[h(X, Y)] \end{aligned}$$

- 따름정리 4.3

$$g(X, Y) = g(X) \text{이고, } h(X, Y) = h(Y) \text{라 하면 } E[g(X) \pm h(Y)] = E[g(X)] \pm E[h(Y)]$$

- 따름정리 4.4

$$g(X, Y) = X \text{이고, } h(X, Y) = Y \text{라 하면 } E[X \pm Y] = E[X] \pm E[Y]$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (8/9)
- 두 확률 변수 X 와 Y 가 서로 독립이면, $E(XY) = E(X)E(Y)$
- 증명

확률변수 $g(x, y)$ 의 수학적 기대값에 대한 정의에 의해서

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y) dx dy$$

X 와 Y 의 주변분포를 각각 $g(x)$ 와 $h(y)$ 라고 하면, X 와 Y 는 독립이므로

$$f(x, y) = g(x)h(y)$$

따라서,

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyg(x)h(y) dx dy = \int_{-\infty}^{\infty} xg(x) dx \int_{-\infty}^{\infty} yh(y) dy = E(X)E(Y)$$

• 따름정리 4.5

확률변수 X 와 Y 가 서로 독립이면 $\sigma_{XY} = 0$

선형결합된 확률변수의 평균과 분산

- 확률변수의 평균을 쉽게 계산할 수 있는 성질 (9/9)

- 예제 4.23

X 와 Y 를 서로 독립이고 다음과 같은 결합확률분포를 가지는 확률변수라 할 때, $E(XY) = E(X)E(Y)$ 가 성립함을 증명하라.

$$f(x, y) = \begin{cases} \frac{x(1 + 3y^2)}{4}, & 0 < x < 2, 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

$$E(XY) = \int_0^1 \int_0^2 xyf(x, y) dx dy = \int_0^1 \int_0^2 \frac{x^2 y(1 + 3y^2)}{4} dx dy = \int_0^1 \left[\frac{x^3 y(1 + 3y^2)}{12} \right]_{x=0}^2 dy = \int_0^1 \frac{2y(1 + 3y^2)}{3} dy = \frac{5}{6}$$

$$E(X) = \int_0^1 \int_0^2 xf(x, y) dx dy = \int_0^1 \int_0^2 \frac{x^2(1 + 3y^2)}{4} dx dy = \int_0^1 \left[\frac{x^3(1 + 3y^2)}{12} \right]_{x=0}^2 dy = \int_0^1 \frac{2(1 + 3y^2)}{3} dy = \frac{4}{3}$$

$$E(Y) = \int_0^1 \int_0^2 yf(x, y) dx dy = \int_0^1 \int_0^2 \frac{xy(1 + 3y^2)}{4} dx dy = \int_0^1 \left[\frac{x^2 y(1 + 3y^2)}{8} \right]_{x=0}^2 dy = \int_0^1 \frac{y(1 + 3y^2)}{2} dy = \frac{5}{8}$$

• 따라서,

$$E(X)E(Y) = \left(\frac{4}{3}\right)\left(\frac{5}{8}\right) = \frac{5}{6} = E(XY)$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 분산을 쉽게 계산할 수 있는 성질 (1/4)

- X 와 Y 가 결합확률분포 $f(x, y)$ 를 가지는 확률변수이고, a, b, c 가 상수일 때, $\sigma_{aX+bY+c}^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$ (1/3)

- 증명

정의에 의하여

$$\sigma_{aX+bY+c}^2 = E\{[(aX + bY + c) - \mu_{aX+bY+c}]^2\}$$

따름정리 4.2와 따름정리 4.4에 의하여

$$\mu_{aX+bY+c} = E(aX + bY + c) = aE(X) + bE(Y) + c = a\mu_X + b\mu_Y + c$$

가 되므로

$$\begin{aligned}\sigma_{aX+bY+c}^2 &= E\{[a(X - \mu_X) + b(Y - \mu_Y)]^2\} \\ &= a^2E[(X - \mu_X)^2] + b^2E[(Y - \mu_Y)^2] + 2abE[(X - \mu_X)(Y - \mu_Y)] \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}\end{aligned}$$

가 된다.

선형결합된 확률변수의 평균과 분산

- 확률변수의 분산을 쉽게 계산할 수 있는 성질 (2/4)

- X 와 Y 가 결합확률분포 $f(x, y)$ 를 가지는 확률변수이고, a, b, c 가 상수일 때, $\sigma_{aX+bY+c}^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$ (2/3)

- 따름정리 4.6

$$b = 0 \text{으로 놓으면 } \sigma_{aX+c}^2 = a^2\sigma_X^2 = a^2\sigma^2$$

- 따름정리 4.7

$$a = 1, b = 0 \text{으로 놓으면 } \sigma_{aX+c}^2 = \sigma_X^2 = \sigma^2$$

- 따름정리 4.8

$$b = 0, c = 0 \text{으로 놓으면 } \sigma_{aX}^2 = a^2\sigma_X^2 = a^2\sigma^2$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 분산을 쉽게 계산할 수 있는 성질 (3/4)

- X 와 Y 가 결합확률분포 $f(x, y)$ 를 가지는 확률변수이고, a, b, c 가 상수일 때, $\sigma_{aX+bY+c}^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$ (3/3)

- 따름정리 4.9

X 와 Y 를 독립인 확률변수라고 하면, $\sigma_{aX+bY}^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$

- 따름정리 4.10

X 와 Y 를 독립인 확률변수라고 하면, $\sigma_{aX-bY}^2 = a^2\sigma_X^2 + b^2\sigma_Y^2$

- 따름정리 4.11

$X_1, X_2, \dots, X_n$ 을 독립인 확률변수라고 하면,

$$\sigma_{a_1X_1+a_2X_2+\dots+a_nX_n}^2 = a_1^2\sigma_{X_1}^2 + a_2^2\sigma_{X_2}^2 + \dots + a_n^2\sigma_{X_n}^2$$

선형결합된 확률변수의 평균과 분산

- 확률변수의 분산을 쉽게 계산할 수 있는 성질 (4/4)
- 예제 4.24

확률변수 X 와 Y 의 분산이 각각 $\sigma_X^2 = 2, \sigma_Y^2 = 4$ 이고, 공분산이 $\sigma_{XY} = -2$ 일 때, 확률변수 $Z = 3X - 4Y + 8$ 의 분산을 구하라.

$$\begin{aligned}\sigma_Z^2 &= \sigma_{3X-4Y+8}^2 = \sigma_{3X-4Y}^2 \\ &= 9\sigma_X^2 + 16\sigma_Y^2 - 24\sigma_{XY} \\ &= (9)(2) + (16)(4) - (24)(-2) = 130\end{aligned}$$

- 예제 4.25

어떤 화학제품의 무더기 속에 섞여 있는 두 종류의 불순물을 X, Y 라고 할 때, X 와 Y 는 서로 독립이고, 분산이 각각 $\sigma_X^2 = 2, \sigma_Y^2 = 3$ 이라고 한다. 확률변수 $Z = 3X - 2Y + 5$ 의 분산을 구하라.

$$\begin{aligned}\sigma_Z^2 &= \sigma_{3X-2Y+5}^2 = \sigma_{3X-2Y}^2 \\ &= 9\sigma_X^2 + 4\sigma_Y^2 \\ &= (9)(2) + (4)(3) = 30\end{aligned}$$

목 차

- 보충
- 확률변수의 평균
- 분산과 공분산
- 선형결합된 확률변수의 평균과 분산
- 체비셰프 정리

체비셰프 정리

- 표준편차 (1/3)

- 정의

- 평균을 기준으로 확률변수의 값들이 얼마나 변동하는 지를 나타내는 척도

- 값의 크기에 따른 해석

- 표준편차가 작은 경우

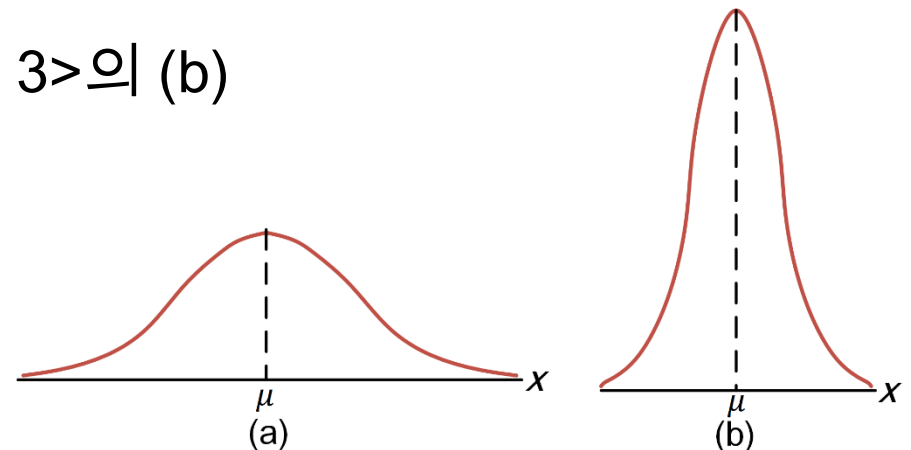
- 대부분의 관측값이 평균(μ) 근처에 몰림
- 평균을 중심으로 한 작은 구간 내에서 값을 취할 가능성이 큼

- 표준편차가 큼

- 관측값이 평균을 중심으로 넓게 퍼져있음
- 평균에서 멀리 떨어진 값을 취할 가능성이 상대적으로 큼

체비셰프 정리

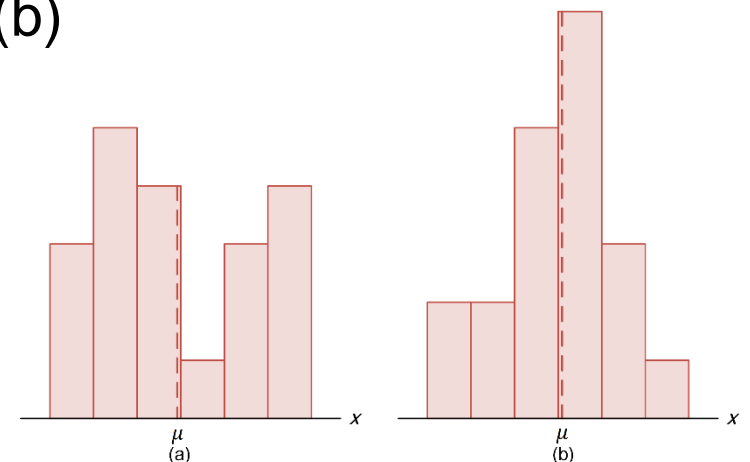
- 표준편차 (2/3)
- 그래프를 통한 해석 (1/2)
 - 연속형 분포
 - 확률의 해석
 - 확률밀도함수 아래의 면적을 통해 해석
 - 표준편차가 큰 경우 - <그림 3>의 (a)
 - 분포 곡선이 넓고 완만함
 - 면적이 넓은 구간에 분포
 - 표준편차가 작은 경우 - <그림 3>의 (b)
 - 분포 곡선이 좁고 뾰족함
 - 면적이 평균 근처에 집중됨



<그림 3> 연속형 관측값들의 산포

체비셰프 정리

- 표준편차 (3/3)
 - 그래프를 통한 해석 (2/3)
 - 이산형 분포
 - 확률의 해석
 - 히스토그램 막대 분포를 통해 해석
 - 표준편차가 큰 경우 - <그림 4>의 (a)
 - 막대가 넓게 흩어져있음
 - 표준편차가 작은 경우 - <그림 4>의 (b)
 - 막대가 평균 근처에 집중됨



<그림 4> 이산형 관측값들의 산포

체비셰프 정리

- 체비셰프 정리(Chebyshev's Inequality) (1/6)

- 개요

- 러시아의 수학자 체비셰프가 평균을 중심으로 대칭인 구간 내 확률이 표준편차와 관련되어 있음을 발견
 - 확률분포곡선 아래 면적이나 히스토그램 면적의 합은 1
 - 특정 두 값 사이의 면적은 확률 변수가 해당 구간의 값을 취할 확률

- 정의

- 확률변수 X 가 평균으로부터 표준편차의 k 배 범위 내의 값을 취할 확률은 적어도 $1 - \frac{1}{k^2}$

- 공식

- $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$

체비셰프 정리

• 체비셰프 정리 (2/6)

• 증명

확률변수 X 의 분산에 대한 정의에 의하여

$$\begin{aligned}\sigma^2 &= E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx \\ &= \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu - k\sigma}^{\mu + k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx \\ &\geq \int_{-\infty}^{\mu - k\sigma} (x - \mu)^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} (x - \mu)^2 f(x) dx\end{aligned}$$

$x \geq \mu + k\sigma$ 이거나 $x \leq \mu - k\sigma$ 이면 $|x - \mu| \geq k\sigma$ 이므로 $(x - \mu)^2 \geq k^2\sigma^2$

따라서,

$$\sigma^2 \geq \int_{-\infty}^{\mu - k\sigma} k^2\sigma^2 f(x) dx + \int_{\mu + k\sigma}^{\infty} k^2\sigma^2 f(x) dx, \int_{-\infty}^{\mu - k\sigma} f(x) dx + \int_{\mu + k\sigma}^{\infty} f(x) dx \leq \frac{1}{k^2}$$

따라서,

$$P(\mu - k\sigma < X < \mu + k\sigma) = \int_{\mu - k\sigma}^{\mu + k\sigma} f(x) dx \geq 1 - \frac{1}{k^2}$$

체비셰프 정리

- 체비셰프 정리 (3/6)
 - 장점
 - 분포 형태를 알지 못해도 사용 가능
 - 평균과 분산만으로 확률 하한 제시 가능
 - 연속형·이산형 확률변수 모두 적용 가능
 - 단점
 - 확률 하한이 실제 확률보다 지나치게 작음
 - 대칭성 등 분포의 구체적 형태 반영 불가

체비셰프 정리

• 체비셰프 정리 (4/6)

• 필요성

- 실제 확률분포의 가정 없이 확률의 하한 제공
 - 평균과 분산만으로 확률변수가 특정 구간에 존재할 확률 하한 제시
 - e.g., 서버 응답시간 분포를 몰라도 평균과 분산을 통해 특정 시간 안에 끝날 확률의 최소값 보장
- 평균 중심의 산포에 대한 확률적 보장 제공
 - 확률변수가 평균을 중심으로 일정 범위 내에 포함될 확률을 분산을 통해 정량적으로 보장
 - e.g., 학생들의 시험 점수가 평균 주변에 얼마나 모여있는지에 대해 분산만으로 $\text{평균} \pm k$ 점 안에 있을 확률의 최소치 확인
- 극단적 분포에 대해서도 성립하는 일반성 보유
 - 분포의 대칭성이나 정규성을 가정하지 않아도 모든 확률분포에 대해 항상 성립하는 보수적 기준을 제공
 - e.g., 소득처럼 일부 초고액 값 때문에 분포가 한쪽으로 치우쳐도 평균과 분산만으로 대부분이 평균 근처에 있을 확률 하한 제시

체비셰프 정리

- 체비셰프 정리 (5/6)

- 예제 4.26

확률변수 X 의 평균은 $\mu = 8$ 이고 분산은 $\sigma^2 = 9$ 이며, 확률분포는 알려져 있지 않다고 할 때 다음을 구하라.

a) $P(-4 < X < 20)$

b) $P(|X - 8| \geq 6)$

- (a) $P(-4 < X < 20)$

$$= P[8 - (4)(3) < X < 8 + (4)(3)] \geq \frac{15}{16}$$

- (b) $P(|X - 8| \geq 6)$

$$= 1 - P(|X - 8| < 6)$$

$$= 1 - P(-6 < X - 8 < 6)$$

$$= 1 - P[8 - (2)(3) < X < 8 + (2)(3)] \leq \frac{1}{4}$$

체비셰프 정리

• 체비셰프 정리 (6/6)

• 예제 4.27

어떤 전자기기의 배터리 수명 X (단위: 시간)에 대해 평균 수명이 10시간, 분산이 4일 때, 해당 배터리 수명이 6시간 이상 14시간 이하일 확률의 최소값을 구하라.

- $P(6 \leq X \leq 14)$
 $= P[10 - (2)(2) < X < 10 + (2)(2)] \geq 1 - \frac{1}{2^2} = \frac{3}{4}$
- $\therefore 75\%$

• 예제 4.28

어떤 웹서비스의 서버 응답 시간 X 를 측정하고 있다.

응답 시간의 정확한 분포 형태는 알 수 없으나, 장기간 로그 분석결과 평균 응답시간은 200ms, 분산은 10,000ms로 알려져 있다.

해당 웹서비스의 서버 응답 시간이 100ms 이상 300ms 이하일 확률의 최소값을 구하라.

- $\mu = 200ms, \quad \sigma^2 = 10000ms, \quad \sigma = 100ms$
- 확률의 범위

$$P(100 \leq X \leq 300) = P(|X - 200| < 100)$$

$$P(|X - 200| < (1)(100)) \geq 1 - \frac{1}{1^2} = 0$$

Thanks!

송 민 희(minhee@pel.sejong.ac.kr)