

확률과 통계

- 정규분포 알아보기 II -

참고서: Head First Statistics

(9장 정규분포 알아보기 II)

작성자 : 이계식

한경대학교, 컴퓨터공학과



정규분포 알아보기 2

1. 주요 내용

정규분포의 연산	정규분포의 활용
* <u>합과 차이</u> 전제 : $X \sim N(\mu_x, \sigma_x^2)$ $Y \sim N(\mu_y, \sigma_y^2)$ • 서로 독립 $\Rightarrow X+Y \sim N(\mu_x+\mu_y, \sigma_x^2 + \sigma_y^2)$ $X-Y \sim N(\mu_x-\mu_y, \sigma_x^2 + \sigma_y^2)$	* <u>이행분포 계산</u> 전제 : $\cdot X \sim B(n, p)$ $\cdot n \cdot p > 5$ $\cdot n \cdot (1-p) > 5$ $\Rightarrow X \sim N(n \cdot p, n \cdot p \cdot q)$ 단, $q = 1 - p$.
* <u>선형변환</u> 전제 : $X \sim N(\mu, \sigma^2)$ $\Rightarrow aX+b \sim N(a\mu+b, a^2\sigma^2)$	* <u>포아송 분포 계산</u> 전제 : $\cdot X \sim P_0(\lambda)$ $\cdot \lambda > 15$ $\Rightarrow X \sim N(\lambda, \lambda)$
* <u>독립관측</u> 전제 : $X \sim N(\mu, \sigma^2)$ • X_1, \dots, X_n 모두 X 의 독립관측 $\Rightarrow X_1 + \dots + X_n \sim N(n \cdot \mu, n \cdot \sigma^2)$	

2. 정규분포들의 합

p. 403

예제: 신장, 신체 체중의 합의 분포

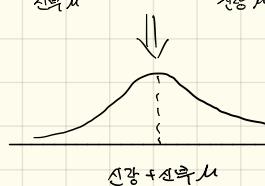
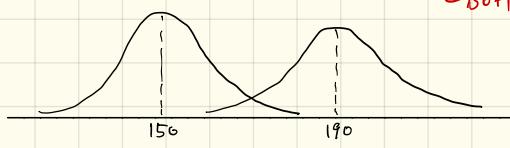
$$\text{전제: 신체의 체중 } X \sim N(150, 400)$$

$$\text{신장의 체중 } Y \sim N(190, 500)$$

$$\Rightarrow \text{신장+신체의 체중} \sim N(340, 900)$$

$$\begin{aligned} & 400 + 500 \\ & \sim 150 + 190 \end{aligned}$$

신장과 신체의
체중은 서로
독립이거나 가정



주의: 크양이 보다
넓적해진다.
이유: 분산이 커진다.

p. 409

설명: 연속데이터를 다루는 두 정규분포의 합도 연속데이터를
다루는 정규분포로 따른다.

또한 서로운 정규분포의 기대치와 분산은 각각의
기대치 또는 출산을 이용하여 아래와 같이 구한다.

$$\text{전제: } X \sim N(\mu_X, \sigma_X^2), \quad Y \sim N(\mu_Y, \sigma_Y^2) \quad (\text{서로 독립})$$

$$\Rightarrow \left\{ \begin{array}{l} E(X+Y) = \mu_X + \mu_Y, \quad \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) \\ X+Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \end{array} \right.$$

p. 411

문제: 신체와 신장의 체중의 합이 380 파운드 이하일 확률은?

$$\begin{aligned} \text{답: } P(X+Y < 380) &= P(Z < \frac{380 - 340}{\sqrt{900}}) \\ &= P(Z < 1.33) \\ &= 0.9082 \end{aligned}$$

3. 정규분포들의 차이

P. 413

예제: 결혼 중매 찰여 남자, 여자의 키 차이

문제: 남자의 키 $X \sim N(71, 20.25)$

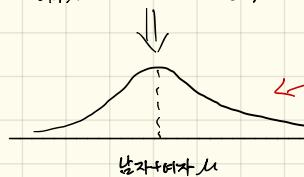
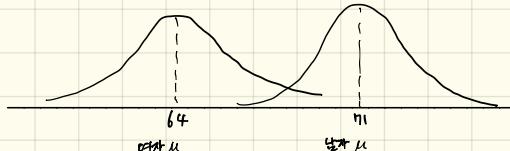
여자의 키 $Y \sim N(64, 16)$

$20.25 + 16$

\Rightarrow 남자의 키와 여자의 키의 차이의 분포 $\sim N(7, 36.25)$

$\frac{7}{71-64}$

남자와 여자의
키는 서로
독립이아니고 가정



주의: 모양이 보다
남작해진다.
이유: 분산이 커진다.

P. 409

설명용 연속데이터를 다루는 두 정규분포의 차이도 연속데이터를
다루는 정규분포로 따른다.

또한 서로운 정규분포의 기대치와 분산은 각각의
기대치 또는 분산을 이용하여 아래와 같이 구한다.

문제: $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ (서로 독립)

$$\Rightarrow \begin{cases} E(X-Y) = \mu_X - \mu_Y, \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) \\ X-Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2) \end{cases}$$

P. 413

문제: 남자와 여자의 키의 차이가 5인치 이상일 확률은?

$$\begin{aligned} \text{답: } P(X-Y > 5) &= P\left(Z > \frac{5-7}{6.02}\right) \\ &= P(Z > -0.33) \\ &= P(Z < 0.33) \\ &= 0.6293 \end{aligned}$$

주의: 정규분포 그래프는
평균을 중심으로해서
좌우 대칭이다.

4. 이항분포를 정규분포로 대체하기

p. 423 예제 : 40명을 무작정 뽑아서 30명이 이상 맞을 확률은?
 \Rightarrow 이항분포 문제임. 즉, $X \sim B(40, 0.25)$ 일.

하지만 40명 중 30명이 이상 맞을 확률은 아래와 같이 계산해야 함:

$$40C_{30} \cdot \left(\frac{1}{4}\right)^{30} \cdot \left(\frac{3}{4}\right)^{10} + 40C_{31} \cdot \left(\frac{1}{4}\right)^{31} \cdot \left(\frac{3}{4}\right)^9 \\ + \dots \\ + 40C_{40} \cdot \left(\frac{1}{4}\right)^{40}$$

{ 그런데 이런 과정은 너무 수고스럽거나, 너무 오래걸리거나,
경우에 따라 계산 불가능.
 ↑ $n!$ 등을 계산해야 하는데 n 이 너무 높을 경우
 컴퓨터를 이용해도 계산이 불가능할 수 있다.

주의: 이 문제의 경우 희귀증후군을 활용할 수 없다. (이유는?)

설득 가능하다 하더라도 별 도움이 되지 않는다.

희귀증후군 역시 많은 경우에 따라 질질적으로 불가능한
 계산이 요구되기 때문이다.

p. 429

여안: 정규분포 활용 가능

설명: $X \sim B(n, p)$ 이고 $n \cdot p > 5$ 와 $n \cdot q > 5$ 라면

$X \sim N(n \cdot p, n \cdot p \cdot q)$ 을 이용해서 이항분포에 대한
 대략적인 값을 구할 수 있다.

5. 정규분포 적용할 때 주의점

이항분포를 정규분포로 대체하여 확률을 계산할 때
주의할 점이 하나 있다.

먼저 아래 예제를 살펴보자.

P.430~431 예제 : 2지선다 12개의 문제 중 5개 이하의 문제를
맞출 확률 계산.

답) $X \sim B(12, \frac{1}{2})$ 가 성립한다.

따라서,

$$\begin{aligned} P(X < 6) &= P(X=0) + \dots + P(X=5) \\ \textcircled{1} \quad \left\{ \right. &= {}^{12}C_0 \cdot (0.5)^{12} + {}^{12}C_1 \cdot (0.5)^{12} + \dots + {}^{12}C_5 \cdot (0.5)^{12} \\ &= 0.389 \end{aligned}$$

반면에 $12 \cdot \frac{1}{2} = 6 > 5$ 가 성립하여 $P(X < 6)$ 을 계산하기 위해
 $X \sim N(6, 3)$ 을 활용할 수 있다. 그런데 아래 계산이
기대하는 다른 결과를 보여준다.

$$\textcircled{2} \quad P(X < 6) = P(Z < \frac{6-6}{\sqrt{3}}) = P(Z < 0) = 0.5$$

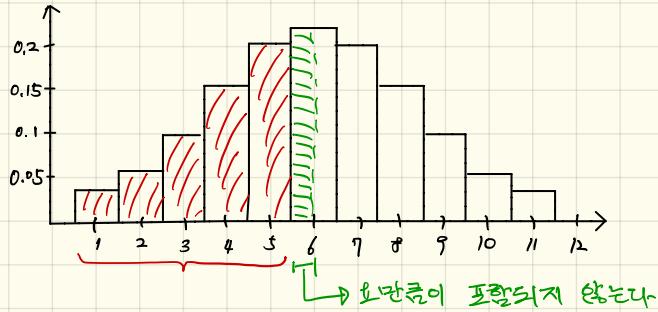
$\textcircled{1}$ 과 $\textcircled{2}$ 의 결과가 다르다!

↑의 경우 정규분포를 사용하여 이항분포의 확률을 계산해도
된다고 하였는데 전혀 비슷하지 않은 결과가 나왔다.
이유는?

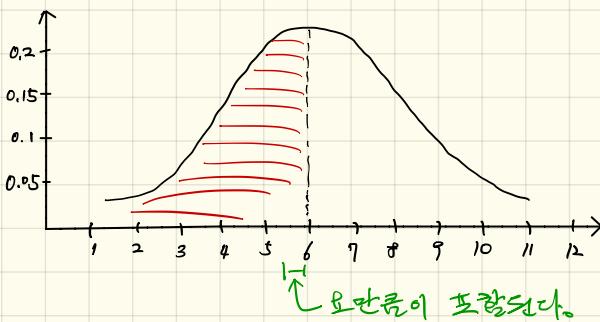
6. 이산 데이타와 정속 데이타의 차이점

P.434

- ① $X \sim B(12, 0.5)$ 일 때 $P(X < 6)$ 을 아래 그래프에서
빨간색으로 표시된 면적에 대응한다.



- ② $X \sim N(6, 3)$ 일 때 $P(X < 6)$ 은 아래 그래프에서
빨간색으로 표시된 면적에 대응한다.



결론: ①과 ②를 비교해 보면 이산 데이타의 경우와
연속 데이타의 경우 확률을 계산할 때 약간의
차이가 발생할 수 있음을 알 수 있다.

7. 연속성 보정

앞서 설명에 빠각 이산데이트를 연속데이트를 이용하여
다치고자 할 때 발생하는 차이를 보정해 주는 과정이
필요하다.

이를 **연속성 보정**이라 부르며, 이산된 값들을 연속값들
으로 옮길 때 반드시 수행해야 하는 간단한 단계조절을
의미한다.

앞에 언급한 예제의 경우, 정규분포 확률을
계산할 때, $P(X < 6)$ 대신에 $P(X < 5.5)$ 를
계산하면 그림과 같은 설명에서 보여준 차이에 대한
보정이 이루어 진다.

실제로, $X \sim N(6, 3)$ 일 때,

$$\begin{aligned} P(X < 5.5) &= P\left(Z < \frac{5.5 - 6}{\sqrt{3}}\right) \\ &= P(Z < -0.29) \\ &= 1 - P(Z < 0.29) \\ &= 1 - 0.6141 \\ &= 0.3859 \end{aligned}$$

이미, 이 값은 이항분포에서 직접 계산한 값인
0.3859와 매우 비슷하다.