

# 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
<p>* <u>합과 차이</u></p> <p>전제 : <math>X \sim N(\mu_x, \sigma_x^2)</math></p> <ul style="list-style-type: none"> <li><math>Y \sim N(\mu_y, \sigma_y^2)</math></li> <li>• 서로 독립</li> </ul> <p><math>\Rightarrow</math></p> <p><math>X+Y \sim N(\mu_x+\mu_y, \sigma_x^2 + \sigma_y^2)</math></p> <p><math>X-Y \sim N(\mu_x-\mu_y, \sigma_x^2 + \sigma_y^2)</math></p>	<p>* <u>이행분포 계산</u></p> <p>전제 : • <math>X \sim B(n, p)</math></p> <ul style="list-style-type: none"> <li><math>n \cdot p &gt; 5</math></li> <li><math>n \cdot (1-p) &gt; 5</math></li> </ul> <p><math>\Rightarrow X \sim N(n \cdot p, n \cdot p \cdot q)</math></p> <p>단, <math>q = 1 - p</math>.</p> <p><u>주의: 연속성 조건 필요</u></p>
<p>* <u>선형변환</u></p> <p>전제 : <math>X \sim N(\mu, \sigma^2)</math></p> <p><math>\Rightarrow</math></p> <p><math>aX+b \sim N(a\mu+b, a^2\sigma^2)</math></p>	<p>* <u>파아송 분포 계산</u></p> <p>전제 : • <math>X \sim P_0(\lambda)</math></p> <ul style="list-style-type: none"> <li><math>\lambda &gt; 15</math></li> </ul> <p><math>\Rightarrow X \sim N(\lambda, \lambda)</math></p> <p><u>주의: 연속성 조건 필요</u></p>
<p>* <u>독립관측</u></p> <p>전제 : <math>X \sim N(\mu, \sigma^2)</math></p> <ul style="list-style-type: none"> <li><math>X_1, \dots, X_n</math> 모두 <math>X</math>의 독립관측</li> </ul> <p><math>\Rightarrow X_1 + \dots + X_n \sim N(n \cdot \mu, n \cdot \sigma^2)</math></p>	

## 2. 정규분포들의 합

p. 403

예제: 신장, 신체 체중의 합의 분포

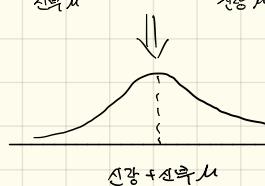
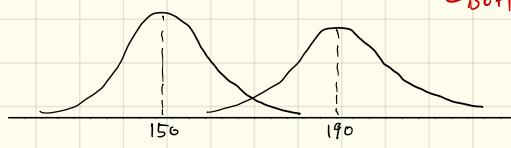
$$\text{전제: 신체의 체중 } X \sim N(150, 400)$$

$$\text{신장의 체중 } Y \sim N(190, 500)$$

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

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

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  $X \sim N(71, 20.25)$

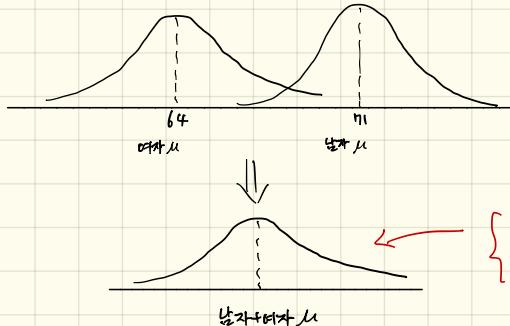
여자의 키  $Y \sim N(64, 16)$

$20.25 + 16$

$\Rightarrow$  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

설명은 연속데이터를 다루는 두 정규분포의 차이도 연속데이터를  
다루는 정규분포를 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)

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

P. 413

문제: 남자와 여자의 키의 차이가 5인치 이상일 확률은?

$$\begin{aligned} \text{답: } P(X-Y > 5) &= P\left(Z > \frac{5-7}{\sqrt{36.25}}\right) \\ &= P(Z > -0.33) \\ &= P(Z < 0.33) \\ &= 0.6293 \end{aligned}$$

주의: 정규분포 그래프는  
평균을 중심으로해서  
좌우 대칭이다.

#### 4. 이항분포를 정규분포로 대체하기

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

하지만 40명 중 30명이 이상 맞출 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.

1-p

## 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,

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

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  
기대와는 다른 결과를 보여준다.

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

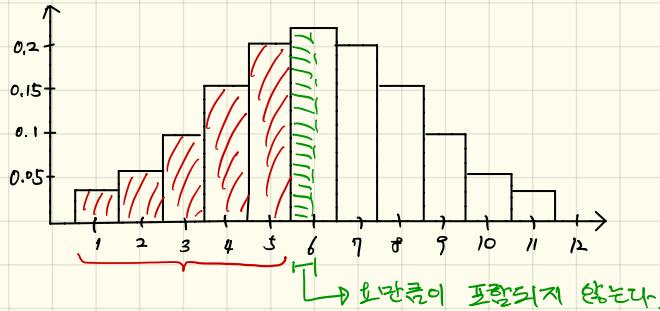
①과 ②의 결과가 다르다!

②의 경우 정규분포를 사용하여 이항분포의 확률을 계산해도  
된다고 하였는데 전혀 비슷하지 않은 결과가 나왔다.  
이유는?

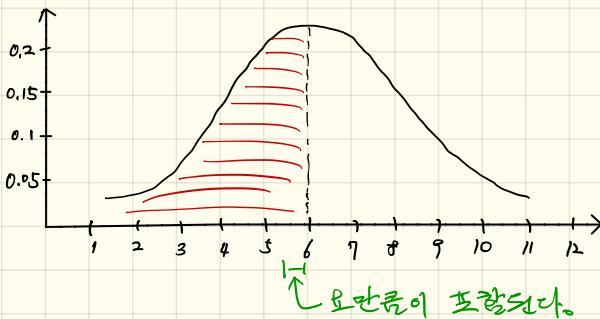
## 6. 이산 데이타와 정속 데이타의 차이점

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.

## 연속성 보정 자세히 보기

P. 437

여제는 2자선과 12문제

① 12문제를 놓아서 9문제 이상 맞출 확률

정답: 이항분포로 계산할 경우에는

$$P(X \geq 8) = P(X \geq 9)$$

$$= {}^{12}C_9 \cdot \left(\frac{1}{2}\right)^9 \cdot \left(\frac{1}{2}\right)^3 + \dots + {}^{12}C_{12} \cdot \left(\frac{1}{2}\right)^{12}$$

를 계산해야 함.

하지만 정답은  $N(6, 3)$ 을 활용할 경우

$\underline{P(X > 8.5)}$ 를 계산해야 함.

$\frac{1}{2}$  연속성 보정!

(여기는 ?)

① 12문제를 놓아서 2개 또 3개 맞출 확률

정답: 이항분포로 계산할 경우에는

$$P(2 \leq X \leq 3) = P(1 < X < 4)$$

$$= {}^{12}C_2 \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^{10} + {}^{12}C_3 \cdot \left(\frac{1}{2}\right)^3 \cdot \left(\frac{1}{2}\right)^9$$

를 계산해야 함.

하지만 정답은  $N(6, 3)$ 을 활용할 경우

$\underline{P(1.5 < X < 3.5)}$ 를 계산해야 함.

$\frac{1}{2}$  연속성 보정!

(여기는 ?)

## 연속성 보정 일반화

P. 437

$X \sim B(n, p)$ 이며  $n \cdot p > 5$ ,  $n \cdot q > 5$  를 성립하면

$X \sim N(np, npq)$ 가 성립한다.

- ① 이항분포에서  $P(X \leq a)$ 를 계산하고자 할 때  
정규분포에서  $P(X < a + 0.5)$ 를 계산한다.
- ② 이항분포에서  $P(X \geq a)$ 를 계산하고자 할 때  
정규분포에서  $P(X > a - 0.5)$ 를 계산한다.
- ③ 이항분포에서  $P(a \leq X \leq b)$ 를 계산하고자 할 때  
정규분포에서  $P(a - 0.5 < X < b + 0.5)$ 를 계산한다.

## 4. 정규분포 선형변환

P.416

$X$ 가 한국인 성인 한 명의 체중이라고 하면 정규분포를 따른다. 체중의 평균이  $\mu$ , 표준은  $\sigma^2$ 이라 하자. 즉, 다음과 성립한다.

$$X \sim N(\mu, \sigma^2)$$

이제, 풍력이 지구의 4배인 행성에 있고 그는 한국인이 그 행성으로 이주했다고 가정하자. 그러면 모든 한국인의 체중이 지구에서의 체중의 4배로 늘어난다. 따라서 체중의 평균도 4배, 즉 4 $\mu$ 가 된다.

반면에 체중의 표준은 지구에서의 경우보다 4 $\sigma$ 이 된다.

따라서 새로운 행성에서의 한국인 체중은 아래의 분포를 따른다:

$$\tilde{X} \sim N(4\mu, 16\sigma^2)$$

→ 사람들의 체중이 4배되었음을 의미함.

결론

기존 데이터의 각각의 값을  
4배한 후 b를 더해주는  
방식을 의미함.

이렇게 기존의 사용된 데이터가  $aX + b$  방식으로 변하면,  
즉, 선형변환을 하면 새로운 데이터의 평균과 표준은 아래와  
같이 바뀐다.

$$aX + b \sim N(a\cdot\mu + b, a^2\sigma^2)$$

## 9. 정규분포의 독립관측

9.4.7

$X$ 가 한국인 성인 한 명의 체중이라고 하면 정규분포를 따른다.  
체중의 평균이  $\mu$ , 표준은  $\sigma$ 이라 하자. 주, 다음이 성립한다.

$$X \sim N(\mu, \sigma^2)$$

이제, 한국인 성인 4명을 무작위로, 그리고 서로 독립적으로  
선별하였을 때 4명 체중의 합의 평균은  $4\cdot\mu$ 이다.

또한 4명 체중의 합의 표준도  $4\cdot\sigma$ 이다.

무작위적이며 서로 독립적인 방식으로 4명을 선별하여  
합한 체중의 분포를  $X+X+X+X$ 로 표기한다.

↑ 4명은 무작위적이며, 서로  
아직 상관없이(즉, 독립적으로)  
선별한다는 의미.

이제,  $X+X+X+X \sim N(4\cdot\mu, 4\cdot\sigma^2)$  이 성립한다.

일반적으로 아래 공식이 성립한다.

$$X_1 + X_2 + \dots + X_n \sim N(n\mu, n\sigma^2)$$

단,  $X_i$ 는 모두  $X$ 의 독립관측이다.

↑  $X$ 를 독립적으로

생각한다는 의미

## 10. 푸아송 분포를 정규분포로 대체하기

9. 447

$X \sim P_0(\lambda)$  이고  $\lambda > 15$  이면  $X \sim N(\lambda, \lambda)$ 을 이용해서  
확률의 근사치를 구할 수 있다.

9. 448

예제 : 라보트레인이라는 플러그스터가 1년에 고장 나는 횟수가  
 $\lambda = 40$  일 때, 1년에 52회 미만으로 고장날  
확률은 얼마인가?

정답 :  $X$  가 1년에 고장나는 횟수를 나타내면  $X \sim P_0(40)$   
이다.  $\lambda > 15$  이므로  $X \sim N(15, 15)$  가  
성립한다.

따라서 푸아송분포에서  $P(X < 52)$ 는 정규분포에서  
 $P(Z < \frac{51.5 - 40}{\sqrt{15}}) = P(Z < 1.82)$   
 $= 0.9656$

이다. 이정분포의 경우처럼 연속성 보정을 사용하는 것에  
주의해야 한다.