확률분포함수

개요

- 확률분포
  - 확률변수가 취할 수 있는 값과 발생할 확률을 대응한 관계
  - 확률변수가 X 가질 수 있는 임의의 실측값 x에 대해

$$F(x) = P(X \leq x)$$

    와 같이 정의된 함수 F를 확률변수 X의 누적분포함수, 또는 간략히 분포함수라고 합니다
  - 분포의 특성인 모수에 따라 분포의 모양이 결정됩니다.
  - 확률질량함수, 확률밀도함수
    - 확률변수 X가 실측값 x를 가질 확률($P(X = x)$)에 대한 함수를 $f(x)$로 나타냅니다.

$$f(x) = P(X = x)$$

    - 확률변수가 취하는 값이 이산형일 경우에는 확률질량함수, 연속형일 경우에는 확률밀도함수라 부릅니다.

베르누이 시행

- 베르누이 시행
  - p의 확률로 원하는 결과가 나타났을 때 '성공'으로, 1-p의 확률로 그렇지 않은 결과가 나타났을 때 '실패'로 하는 두 가지 결과가 나타나는 확률실험입니다.
  - 성공 확률 p가 베르누이 시행의 모수입니다.
  - 확률변수 X가 베르누이 시행에 따라 성공일 때 1, 실패일 때 0을 가질 경우 확률질량함수는 다음과 같습니다.

$$f(x) = p^x \cdot (1-p)^{1-x}, \qquad x = \begin{cases} \text{성공} & 1 \\ \text{실패} & 0 \end{cases}$$

  - 예) 주사위를 던져 3의 배수의 눈이 나오면 상금얻는 게임
    - 성공 : 3의 눈, 6의 눈이 나오는 경우, X=1

$$P(X=1) = p^{x=1} \cdot (1-p)^{1-(x=1)} = p$$

    - 실패 : 성공의 경우가 아닌 눈이 나오는 경우, X=0

$$P(X=0) = p^{x=0} \cdot (1-p)^{1-(x=0)} = 1-p$$

베르누이 시행

- 기댓값과 분산

  - 기댓값 : p

$$E(X) = \sum_{\text{모든 } x} x \cdot P(X = x)$$

$$= \sum_{\text{모든 } x} x \cdot f(x) = 0 \cdot (p^0 \cdot (1-p)^1) + 1 \cdot (p^1 \cdot (1-p)^0) = p$$

  - 분산 : $p \cdot (1-p)$

$$Var(X) = E(X^2) - (EX)^2$$

$$= \sum_{\text{모든 } x} \{x^2 \cdot f(x)\} - p^2$$

$$= \sum_{\text{모든 } x} \{(0^2 \cdot (p^0 \cdot (1-p)^1) + 1^2 \cdot (p^1 \cdot (1-p)^0)\} - p^2$$

$$= p - p^2 = p(1-p)$$

이항분포

- 개요
  - 성공 확률이 p로 동일한 베르누이 시행을 n번 반복해서 실험하는 경우
    - 실험이 n번 반복되더라도 성공 확률 p는 변하지 않고 동일
    - 각 실험이 서로 독립적으로 시행 (iid)
  - n번 반복 실험에서 성공의 횟수가 따르는 분포를 이항분포라고 합니다.
  - 이항분포의 모수
    - n : 시행의 횟수
    - p : 성공의 확률
  - 이항분포의 표기 : 위의 두 모수를 이용하여 $B(n,p)$
    - 확률변수 X가 이항분포를 따를 때 $X \sim B(n,p)$와 같이 나타냅니다.

이항분포

- 이항분포의 확률질량함수
  - 주사위를 굴려 3의 배수가 나올 때를 성공으로 하는 실험(성공의 확률이 1/3인 베르누이 시행)을 3번 독립으로 반복해서 실험할 때의 성공 횟수의 확률질량함수를 구해 봅시다.
    - 확률변수 X를 성공의 확률이 1/3인 베르누이 시행을 3번 독립으로 반복해서 실험하였을 때 성공의 횟수라 할 때 확률변수 X는 다음과 같은 이항분포를 따릅니다.

$$X \sim B\left(n = 3, p = \frac{1}{3}\right)$$

    - 확률변수 X가 가질 수 있는 값은 {0, 1, 2, 3}이고 각각의 확률들을 구해봅시다.
      - x_i는 i번째 주사위를 굴렸을 때 성공과 실패를 각각 1과 0을 가짐을 나타냅니다.

이항분포

- X=0 일 때 : 성공의 횟수가 0일때
  - $x_1 = 0, x_2 = 0, x_3 = 0$일 때의 베르누이 시행의 확률질량함수($f_{Ber}(x_i = 0)$)를 구합니다. (성공의 확률 p=1/3)
    - $f_{Ber}(x_1 = 0) = \frac{1}{3}^{x_1=0} \frac{2}{3}^{1-(x_1=0)}$
    - $f_{Ber}(x_2 = 0) = \frac{1}{3}^{x_2=0} \frac{2}{3}^{1-(x_2=0)}$
    - $f_{Ber}(x_3 = 0) = \frac{1}{3}^{x_3=0} \frac{2}{3}^{1-(x_3=0)}$
  - 모두 실패한 경우는 $x_1 = 0, x_2 = 0, x_3 = 0$의 한가지 경우만 있습니다.
  - 각 실험이 독립이므로 각 확률들의 곱으로 나타낼 수 있습니다.
    - $f_{Ber}(x_1 = 0) \cdot f_{Ber}(x_2 = 0) \cdot f_{Ber}(x_3 = 0)$
    - $P(X = 0) = \frac{1}{3}^{x_1=0} \frac{2}{3}^{1-(x_1=0)} \cdot \frac{1}{3}^{x_2=0} \frac{2}{3}^{1-(x_2=0)} \cdot \frac{1}{3}^{x_3=0} \frac{2}{3}^{1-(x_3=0)} = \frac{1}{3}^{0} \frac{2}{3}^{3}$

이항분포

- X=1 일 때 : 성공의 횟수가 1일때
  - x_1, x_2, x_3 중 한 경우는 1을 갖고(성공) 나머지 두 경우는 0을 가지는(실패) 경우로 다음과 같이 세 가지 상태를 갖습니다.
  - 첫 번째 성공했을 때 : $x_1 = 1, x_2 = 0, x_3 = 0$
    - $\frac{1}{3}^{x_1=1}\ \frac{2}{3}^{1-(x_1=1)} \cdot \frac{1}{3}^{x_2=0}\ \frac{2}{3}^{1-(x_2=0)} \cdot \frac{1}{3}^{x_3=0}\ \frac{2}{3}^{1-(x_3=0)} = \frac{1}{3}^{1}\ \frac{2}{3}^{2}$
  - 두 번째 성공했을 때 : $x_1 = 0, x_2 = 1, x_3 = 0$
    - $\frac{1}{3}^{x_1=0}\ \frac{2}{3}^{1-(x_1=0)} \cdot \frac{1}{3}^{x_2=1}\ \frac{2}{3}^{1-(x_2=1)} \cdot \frac{1}{3}^{x_3=0}\ \frac{2}{3}^{1-(x_3=0)} = \frac{1}{3}^{1}\ \frac{2}{3}^{2}$
  - 세 번째 성공했을 때 : $x_1 = 0, x_2 = 0, x_3 = 1$
    - $\frac{1}{3}^{x_1=0}\ \frac{2}{3}^{1-(x_1=0)} \cdot \frac{1}{3}^{x_2=0}\ \frac{2}{3}^{1-(x_2=0)} \cdot \frac{1}{3}^{x_3=1}\ \frac{2}{3}^{1-(x_3=1)} = \frac{1}{3}^{1}\ \frac{2}{3}^{2}$
  - 세가지 경우의 $\frac{1}{3}^{1}\ \frac{2}{3}^{2}$ 이 있는 것으로 $P(X = 1) = 3 \cdot \frac{1}{3}^{1}\ \frac{2}{3}^{2}$ 입니다.

이항분포

- X=2 일 때 : 성공의 횟수가 2일때
  - x_1, x_2, x_3 중 두 경우는 1을 갖고(성공) 나머지 한 경우는 0을 가지는(실패) 경우로 다음과 같이 세 가지 상태를 갖습니다.
  - 첫 번째와 두 번째가 성공했을 때 : $x_1 = 1, x_2 = 1, x_3 = 0$
    - $\frac{1}{3}^{x_1=1}\ \frac{2}{3}^{1-(x_1=1)} \cdot \frac{1}{3}^{x_2=1}\ \frac{2}{3}^{1-(x_2=1)} \cdot \frac{1}{3}^{x_3=0}\ \frac{2}{3}^{1-(x_3=0)} = \frac{1}{3}^{2}\ \frac{2}{3}^{1}$
  - 두 번째와 세 번째가 성공했을 때 : $x_1 = 0, x_2 = 1, x_3 = 1$
    - $\frac{1}{3}^{x_1=0}\ \frac{2}{3}^{1-(x_1=0)} \cdot \frac{1}{3}^{x_2=1}\ \frac{2}{3}^{1-(x_2=1)} \cdot \frac{1}{3}^{x_3=1}\ \frac{2}{3}^{1-(x_3=1)} = \frac{1}{3}^{2}\ \frac{2}{3}^{1}$
  - 첫 번째와 세 번째가 성공했을 때 : $x_1 = 1, x_2 = 0, x_3 = 1$
    - $\frac{1}{3}^{x_1=1}\ \frac{2}{3}^{1-(x_1=1)} \cdot \frac{1}{3}^{x_2=0}\ \frac{2}{3}^{1-(x_2=0)} \cdot \frac{1}{3}^{x_3=1}\ \frac{2}{3}^{1-(x_3=1)} = \frac{1}{3}^{2}\ \frac{2}{3}^{1}$
  - X=1일때와 마찬가지로 세가지 경우가 있어, $P(X = 2) = 3 \cdot \frac{1}{3}^{2}\ \frac{2}{3}^{1}$ 입니다.

이항분포

- X=3 일 때 : 성공의 횟수가 3일때
  - $x_1 = 1, x_2 = 1, x_3 = 1$ 인, 즉 모두 성공한 경우로 다음과 같은 한가지 경우입니다.
  - 즉, $\frac{1}{3}^{x_1=1}\frac{2}{3}^{1-(x_1=1)} \cdot \frac{1}{3}^{x_2=1}\frac{2}{3}^{1-(x_2=1)} \cdot \frac{1}{3}^{x_3=1}\frac{2}{3}^{1-(x_3=1)} = \frac{1}{3}^{3}\frac{2}{3}^{0}$
- 모두 모아 봅시다.

$$P(X=x) = f(x) = \begin{cases} X=0, & 1 \cdot \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^3 = \frac{1 \cdot 8}{27} \\ X=1, & 3 \cdot \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^2 = \frac{3 \cdot 4}{27} \\ X=2, & 3 \cdot \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^1 = \frac{3 \cdot 2}{27} \\ X=3, & 1 \cdot \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^0 = \frac{1 \cdot 1}{27} \end{cases}$$

이항계수 $= \binom{n}{x} = \frac{n!}{(n-x)!x!}$ 확률 $p^x(1-p)^{n-x}$

이항분포

- 이상으로 부터 확률변수 X가 이항분포를 따를 때의 확률질량함수는 다음과 같습니다.
  - $P(X = x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \dots, n$
- 이항분포의 기댓값과 분산
  - 앞서 사용한 주사위를 굴려 3의 배수가 나오는 경우의 성공 횟수에 대한 분포($B(n = 3, p = \frac{1}{3})$)를 이용하여 기댓값의 정의에 따라 확률변수 X가 이항분포를 따를 때의 기댓값을 구해 봅시다.

$$E(X) = \sum_{\text{모든 } x} x \cdot P(X = x) = \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x}$$

이항분포

- $X \sim B\left(3, \frac{1}{3}\right)$ 이고, $E(X) = \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x}$ 이므로,

$$E(X) = \sum_{x=0}^{n} x \cdot \binom{n}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{n-x}$$

$$= \left(0 \times 1 \cdot \left(\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^3\right) + \left(1 \times 3 \cdot \left(\frac{1}{3}\right)^1 \left(\frac{2}{3}\right)^2\right) + \left(2 \times 3 \cdot \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right)^1\right) + \left(3 \times 1 \cdot \left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^0\right)$$

$$= 0 \times \frac{8}{27} + 1 \times \frac{12}{27} + 2 \times \frac{6}{27} + 3 \times \frac{1}{27} = \frac{0+12++13+4}{27} = \frac{27}{27} = 1$$

  - 여기서 구한 값인 1은 이항분포의 두 모수 n=3과 p=1/3 의 곱 np와 같습니다.
- 이항분포의 기댓값$(E(X))$: np

이항분포

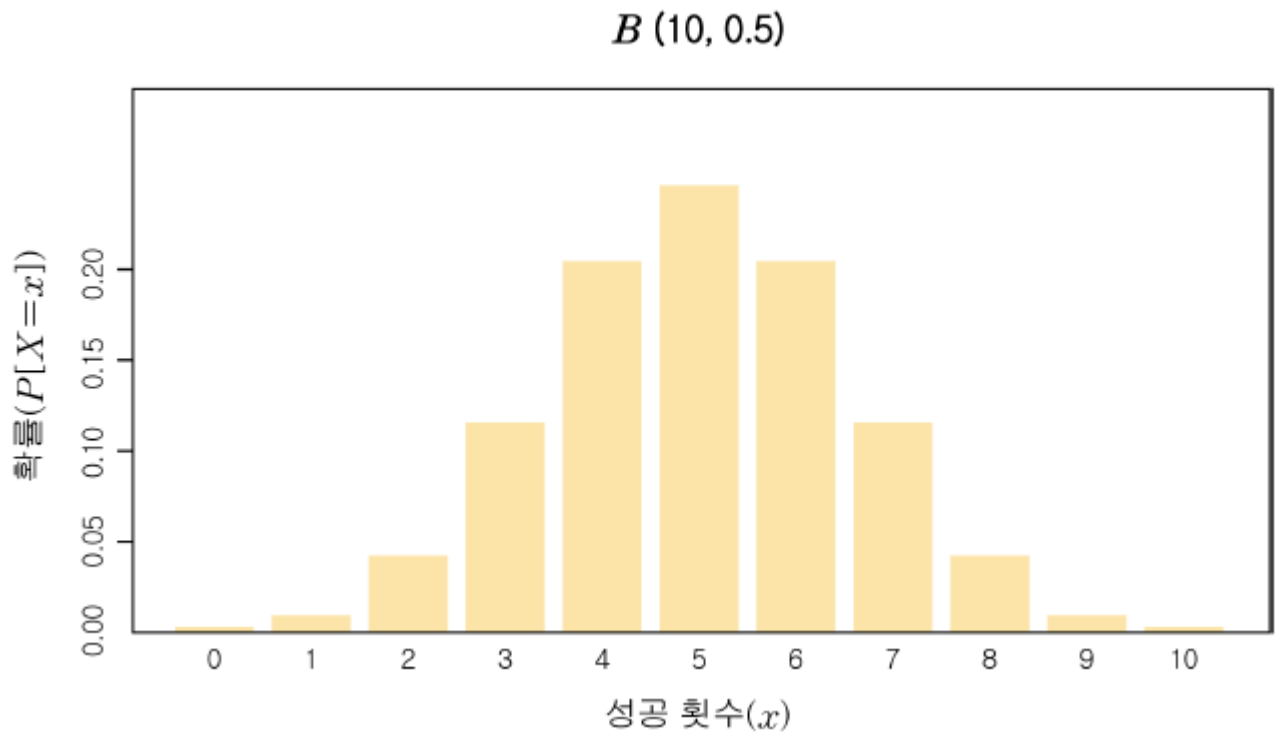
- 이항분포의 분산
  - 확률분수의 분산을 구하는 간편식을 이용하여 (교재에 잘못된 표현) 구할 수 있으며 그 값은 $np(1-p)$로 알려져 있습니다.
  - 다음 식의 전개 과정은 QR 코드(위키피디어)를 통해 확인할 수 있습니다.

$$Var(X) = E(X^2) - \big(E(X)\big)^2 = \sum_{\text{모든 } x} x^2 \cdot P(X = x) - \big(E(X)\big)^2$$

$$= \sum_{x=0}^{n} x^2 \cdot \binom{n}{x} p^x (1-p)^{n-x} - (np)^2 \qquad \text{(식 3.26)}$$

이항분포

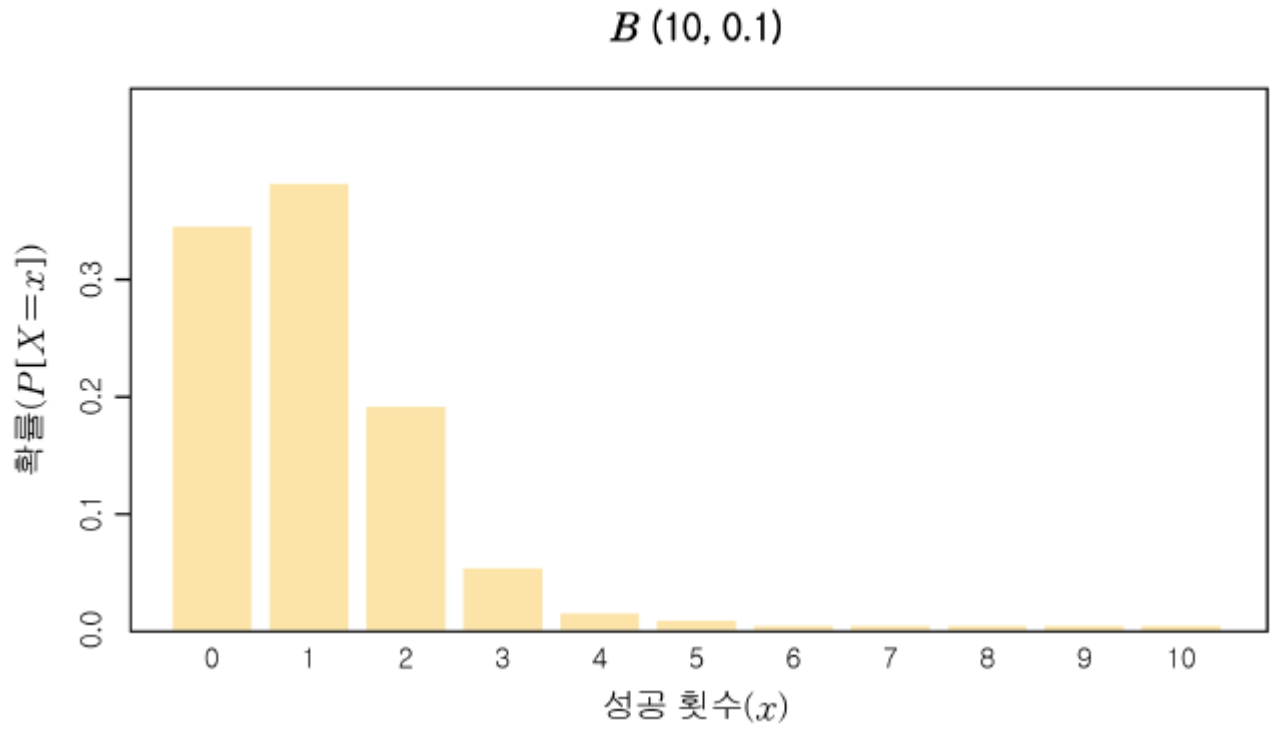
- 성공 확률 p의 변화에 따른 이항분포의 모양 변화
  - 어떤 확률변수 X가 시행 횟수가 10, 성공 확률이 0.5인 이항분포를 따른다고 할 때($X \sim B(10, 0.5)$), 확률변수 X의 분포도는 다음과 같습니다.



- 기

이항분포

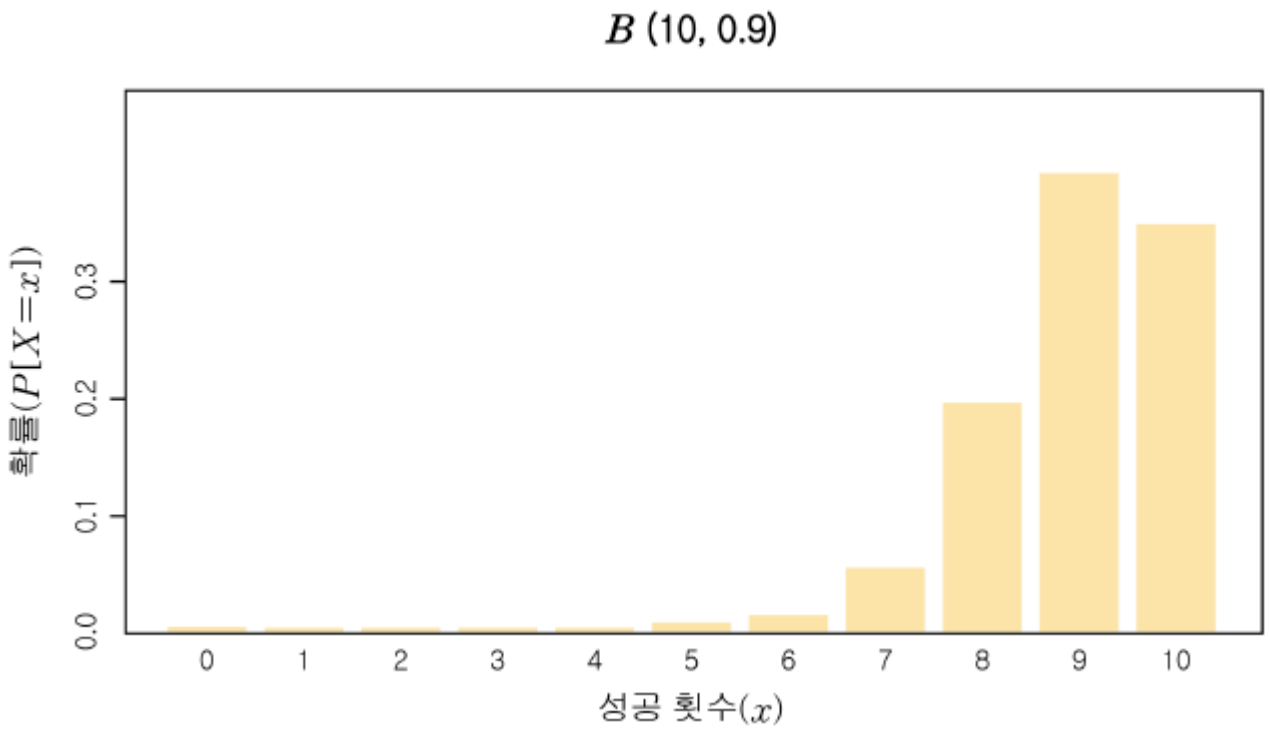
- 성공의 확률을 0.1인 경우에는 다음 그림과 같습니다.



- 기댓값인 1이 최빈값이며 꼬리가 오른쪽으로 길어집니다.
- 성공의 확률이 0.5보다 작은 경우 이러한 모양을 갖습니다.

이항분포

- 성공의 확률이 0.9인 경우에는 다음 그림과 같습니다.



- 기댓값인 9가 최빈값이며 꼬리가 왼쪽으로 길어집니다.
- 성공의 확률이 0.5보다 큰 경우 이러한 모양을 갖습니다.

이항분포

- 개요
  - 확률변수 X가 시행의 횟수가 6이고 성공의 확률이 1/3인 이항분포를 따를 때 R이 내장하고 있는 이항분포와 관련된 함수를 이용하여 각종 확률을 계산해 봅시다.
- Step #1) 이항분포 함수 사용을 위한 모수 준비

```
① n <- 6
② p <- 1/3
③ x <- 0:n
```

  - 1줄 : 시행 횟수는 6이고 이를 변수 n에 저장합니다.
  - 2줄 : 성공 확률은 1/3이고 이를 변수 p에 저장합니다.
  - 3줄 : 확률변수 X가 가질 수 있는 값은 {0, 1, 2, 3, 4, 5, 6} 으로 이를 벡터로 변수 x에 저장합니다.

이항분포

- Step #2) 확률질량함수($P(X = x)$) 를 구합니다.
  - 이항분포의 확률질량함수를 구하는 R 함수인 dbinom()을 사용합니다.

```
dbinom(x, size, prob)
- x : 이항분포의 성공의 횟수의 벡터(원소 1개짜리 포함)
- size : 시행의 횟수
- prob : 성공의 확률
```

```
⑤ ( dbinom(2, size=n, prob=p) )
⑥ ( dbinom(4, size=n, prob=p) )
⑦ ( px <- dbinom(x, size=n, prob=p) )
⑧ plot(x, px, type="s", xlab="성공 횟수(x)", ylab="확률
   (P[X=x])", main="B(6, 1/3)")
```

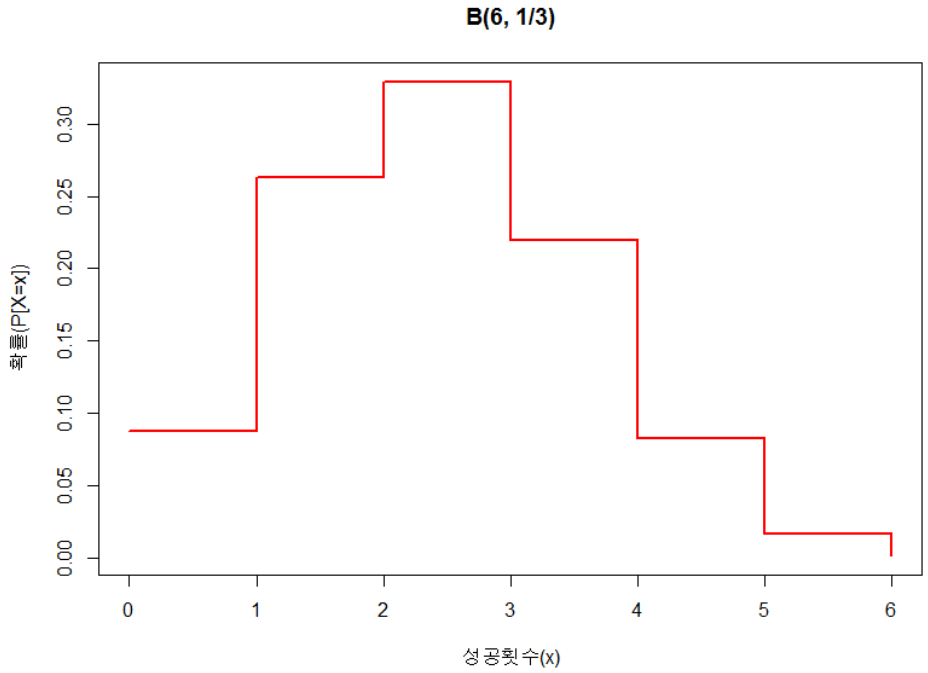
이항분포

- 5, 6줄 : 확률변수 X가 2와 4를 가질 확률을 각각 계산하고 출력합니다. R에서 변수 할당문의 앞과 뒤를 괄호( )로 둘러싸면 바로 결과를 출력해줍니다.
- 7줄 : x를 통해 전달된 성공의 횟수별 확률을 계산한 벡터를 반환합니다.
  - 코드에서는 x가 확률변수 X가 가질 수 있는 값인 0부터 6을 저장한 벡터로, 각각의 확률을 계산한 결과를 변수 px에 저장하고 출력합니다.

```
> (dbinom(2, size=n, prob=p))
[1] 0.3292181
> (dbinom(4, size=n, prob=p))
[1] 0.08230453
> (px <- dbinom(x, size=n, prob=p))
[1] 0.087791495 0.263374486 0.329218107 0.219478738
[5] 0.082304527 0.016460905 0.001371742
```

이항분포

- 8줄 : 위에서 계산한 벡터 x를 이용하여 이항분포 그래프를 작성합니다. plot() 함수의 전달인자로 type에 “s”(소문자)를 사용할 경우, 시작 값을 수평으로 먼저 그리는 계단 형태의 그래프를 작성합니다.
  - type으로 “s”를 가질 때의 plot() 함수의 모양을 설명하기 위한 예제로 이산형 확률분포에는 각 값들이 떨어져 있는 그래프가 적절합니다. (type에 “h”를 넣어보세요)



이항분포

- Step #3) 누적분포함수($\boldsymbol{P(X \leq x)}$) 를 구합니다.
  - 이항분포의 누적분포함수를 구하는 R 함수인 pbinom()을 사용합니다.

```
pbinom(x, size, prob)
- x : 이항분포의 성공의 횟수의 벡터(원소 1개짜리 포함)
- size : 시행의 횟수
- prob : 성공의 확률
```

```
10.(pbinom(2, size=n, prob=p))
11.(pbinom(4, size=n, prob=p))
12.(pbinom(4, size=n, prob=p) - pbinom(2, size=n,
   prob=p))
```

이항분포

- 10줄 : 성공 횟수가 2 이하일 확률을 구하고 출력합니다.
- 11줄 : 성공 횟수가 4 이하일 확률을 구하고 출력합니다.
- 12줄 : (성공 횟수가 4 이하일 확률) - (성공 횟수가 2 이하일 확률)
  - 성공 횟수가 4 이하 2 초과(3 이상)일 확률($P(2 < X \leq 4)$)을 구합니다.

```
> (pbinom(2, size=n, prob=p))
[1] 0.6803841
> (pbinom(4, size=n, prob=p))
[1] 0.9821674
> (pbinom(4, size=n, prob=p) - pbinom(2, size=n, prob=p))
[1] 0.3017833
```

이항분포

- Step #4) 분위 p에 해당하는 확률변수 X의 값 x를($\boldsymbol{P(X \leq x) = p}$) 를 구합니다.
  - 이항분포의 분위 p에 해당하는 값을 찾는 R 함수인 qbinom()을 사용합니다.

```
qbinom(p, size, prob)
- p : 알고자 하는 분위 벡터(원소 1개짜리 포함)
- size : 시행의 횟수
- prob : 성공의 확률
```

```
14. (qbinom(0.1, size=n, prob=p))
15. (qbinom(0.5, size=n, prob=p))
```

이항분포

- 14, 15줄 : 확률변수 X가 $B(6, \frac{1}{3})$을 따를때, 확률변수 X의 확률분포에서 10%(0.1)와 50%(중앙값, 0.5)에 해당하는 x의 값을 출력합니다.
  - 이산형 분포함수에서는 x가 서로 떨어져 있으므로 누적확률값이 전달하는 분위를 포함하는 가장 작은 값이 출력됩니다.
  - p 가 0.1인 경우 X가 0이하일 확률이 0.09, 1 이하일 확률이 0.35, 2이하일 확률이 0.68로, 이 중 0은 해당하지 않고 0.1 이상의 확률값을 갖는 x 중({1, 2, 3, ...}) 가장 작은 값인 1이 이에 해당합니다.

이항분포

- Step #4) 이항분포를 따르는 모집단으로부터 n개의 표본 추출
  - 이항분포로부터 난수를 생성하는 R 함수인 rbinom()을 사용합니다.

```
rbinom(n, size, prob)
- n : 생성하고자 하는 난수의 개수
- size : 시행의 횟수
- prob : 성공의 확률
```

  - 17줄 : $B\left(6, \frac{1}{3}\right)$ 인 모집단으로부터 10개의 확률표본을 추출합니다
    - 난수는 실행할 때마다 다르게 나타납니다.

```
17. rbinom(10, size=n, prob=p)
```

```
> (rbinom(10, size=n, prob=p))
[1] 4 1 2 2 3 1 1 3 2 3
```

이항분포

- 개요
  - 기댓값과 분산을 구하는 식을 이용하여 확률변수 X가 이항분포 $B(6,\frac{1}{3})$를 따를 때의 기댓값과 분산을 R에서 구해봅시다.
- Step #1) 기댓값과 분산을 위한 정보 생성

```
① n <- 6
② p <- 1/3
③ x <- 0:n
④ px <- dbinom(x, size=n, prob=p)
```

  - 1~3줄 : 앞서 사용한 것과 같이 시행의 횟수(n), 성공의 확률(p), 확률변수가 가질 수 있는 모든 값(x)를 설정합니다.
  - 4줄 : 확률변수 X가 가질 수 있는 값인 0부터 6에 대해 dbinom() 함수를 이용해 각각의 확률을 계산하고, 변수 px에 저장합니다.

이항분포

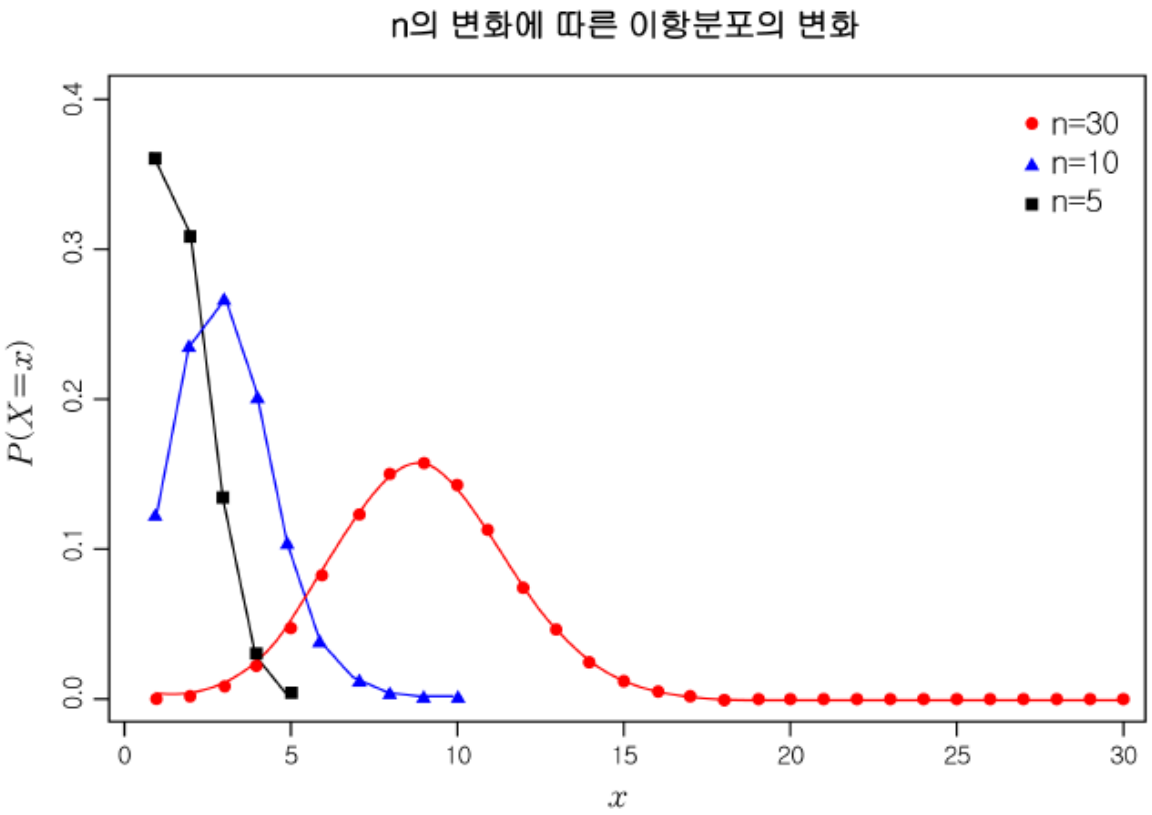
- Step #2) 기댓값과 분산의 계산식을 이용하여 R로 구합니다.
  - $E(X) = \sum_{\text{모든 } x} x \cdot P(X = x)$
  - $Var(X) = E(X^2) - \big(E(X)\big)^2 = \sum_{\text{모든 } x} x^2 \cdot P(X = x) - \big(E(X)\big)^2$

```
⑤  (ex <- sum( x * px ))
⑥  ex2 <- sum( x^2 * px )
⑦  (varx <- ex2 - ex^2)
```

  - 5줄 : 확률변수의 기댓값의 정의를 이용하여 기댓값을 구합니다.
    - x는 확률변수 X가 가질 수 있는 값들이 저장된 벡터
    - px는 확률변수 X가 실측값 x를 가질 때의 이항분포의 확률값이 저장된 벡터
    - 두 벡터를 곱하고 sum() 함수를 이용하여 이를 더합니다.
  - 6, 7줄 : 분산의 간편식을 이용하여 분산을 구합니다.
    - 6줄($E(X^2)$) : x를 제곱하고 이를 px와 곱한후 모두 더한 값을 ex2에 저장합니다.
    - 7줄 : 6줄에서 구한 $E(X^2)$에서 5줄에서 구한 $E(X)$의 제곱값을 빼고 varx에 저장하고 출력합니다.

정규분포

- 개요
  - p가 0.3일 때 시행횟수 n의 변화에 따른 이항분포의 변화
    - 다음 그림은 p가 0.3으로 고정하고 n=5, n=10, n=30 일 때의 이항분포입니다.

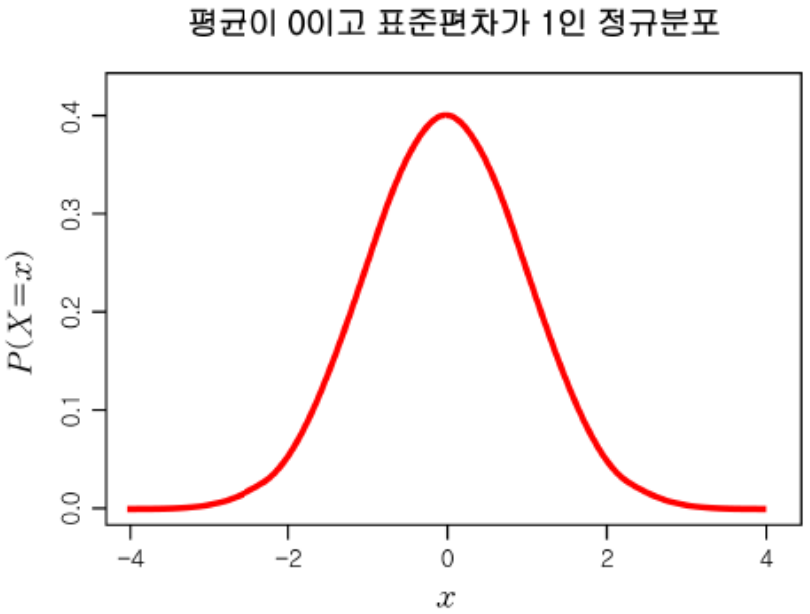


정규분포

- 이항분포에서 시행 횟수 n이 커지면, 그에 따라 이를 따르는 확률변수 X가 갖는 확률($P(X = x)$) 계산은 복잡해집니다.
- 프랑스 태생의 수학자 드무아브르(1667~1754) 가 성공 확률이 0.5이고 시행 횟수 n이 아주 큰 이항분포가 어떤 함수와 비슷해지는 것을 발견하였습니다.
  - 이 함수의 모양은 앞선 그림에서 n이 30인 경우와 많이 닮았습니다.
  - 좌우가 대칭인 종모양(확률분포의 확률값이 x축에 가까이 다가가나 확률이 0이되지않는)의 형태와 유사합니다.
  - n 이 충분히 크다면 이산형이 아닌 연속형처럼 다루는 것이 가능합니다.
  - 이런 형태를 갖는 분포는, 이항분포가 아닌 다른 분포에서도 이와 닮아감을 밝혔습니다. (라플라스(1749~1827))
  - 관측 오차가 이러한 분포를 따른다는 점이 발견되어 폭넓게 사용되었습니다.(가우스(1777~1855))

정규분포

- 정규분포
  - ① 종모양의 형태를 가집니다.
    - 양 끝이 아주 느린 속도로 감소하지만, 축에 닿지 않고 -∞와 ∞까지 계속됩니다.
  - ② 평균을 중심으로 좌우대칭입니다
  - ③ 평균 주변에 많이 몰려 있으며 양 끝으로 갈수록 줄어듭니다.
  - ④ 평균과 표준편차로 분포의 모양을 결정합니다.
    - 정규분포의 모수는 평균 μ와 표준편차 σ(분산 σ^2)로, $N(\mu, \sigma^2)$으로 나타냅니다.
- 정규분포의 확률밀도함수
  - $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty \le x \le \infty$



정규분포

- 표준정규분포
  - 평균이 0이고 표준편차가 1인 정규분포($N(0,1^2)$) 를 **표준정규분포**라하고 대문자 Z로 표시합니다.
  - 모든 정규분포는 표준정규분포로 변환할 수 있습니다.
    - 확률변수 X가 평균 μ와 표준편차 σ인 정규분포를 따른다고 할 때,

$$Z = \frac{X - \mu}{\sigma}, \qquad Z \sim N(0,1^2)$$

  - 표준정규분포를 사용하면 보다 손쉬운 계산이 가능합니다
    - 예 : 어느 대학교 남학생들 키의 평균은 170cm, 표준편차는 6cm입니다. 이 대학교에서 남학생의 키가 182cm 이상일 확률은 다음과 같이 구합니다.(남학생의 키는 정규분포를 따로 있는 것으로 가정합니다.)

$$P(X \geq 182) = 1 - P(X \leq 182) = 1 - \int_{-\infty}^{182} \frac{1}{6\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-170}{6}\right)^2} dt$$

정규분포

  - 수도 없이 많은 정규분포별로 이런 계산을 해야만 했습니다.
  - 만일, 어느 한 정규분포로 각 확률 값들을 계산해 놓고 다른 정규분포들을 이 분포로의 변환이 가능하며 필요에 따라 다시 또 원래대로 돌아갈 수 있게 한다면 참으로 편리할 것입니다.
- 평균이 0이고 표준편차가 1인 표준정규분포로 각 값을 계산해 표로 만들고, 다음과 같은 과정을 통해 그 값을 구해보았습니다.
  ① 임의의 정규분포를 표준정규분포로 변환합니다.
  ② 구하고자 하는 값을 미리 계산된 표준정규분포의 분포표를 통해 구합니다(부록 C).
  ③ 구한 값을 원래의 정규분포로 변환합니다.

정규분포

- 표준화 변환을 통한 표준정규분포로 계산
  - $z = \frac{x-\mu}{\sigma} = \frac{182-170}{6} = \frac{12}{6} = 2$
  - 이를 이용하여 표준정규분포에서 구하면 다음과 같습니다.
    - $P(Z \geq 2) = 1 - P(Z \leq 2)$
    - [부록 C]의 표준정규분포표에서 z값이 2가 되는 값, 즉 행에서 2.0을 찾고 열에서 0.00을 찾은 값은 0.977(유효숫자 세째자리)입니다. 표로부터 표준정규분포에서 2보다 작을 확률은 0.977이고, z가 2보다 클 확률은 1-0.977≈0.023입니다.
- 이제 다시 원래의 정규분포로 돌아가서 z 값으로 변환하여 2가 된 원래의 값을 구해보면 182입니다. 이를 통해 182cm보다 클 확률은 0.023이 됨을 알 수 있습니다.

정규분포

- 개요
  - 표준정규분포를 이용하여 분포표를 읽어 정규분포를 계산할 수 있습니다.
  - 우리는 R을 이용하여 정규분포와 관련된 각종 계산과 그래프를 그려봅시다.
  - 확률변수 X가 평균이 170이고 표준편차가 6인 정규분포($N(170, 6^2)$)를 따를 때로 각종 값들을 계산해 봅시다.
- Step #1) 각종 모수 및 그래프를 위한 작업

```
1. options(digits=3)
2. mu <- 170
3. sigma <- 6
4. ll <- mu - 3*sigma
5. ul <- mu + 3*sigma
```

정규분포

- 1줄 : 출력물이 세 자릿수가 되도록 합니다.
- 2줄 : 평균은 170이고, 이를 변수 mu에 저장합니다.
- 3줄 : 표준편차는 6이고, 이를 변수 sigma에 저장합니다.
- 4, 5줄 : 정규분포에서 확률변수가 가질 수 있는 값의 범위가 $-\infty \leq X \leq \infty$로 그래프를 그리기에 너무 넓습니다.
  - 전체 구간보다는 평균 중심으로 세 배의 표준편차 범위로 한정해서 구하려고 합니다.
  - 변수 ll에는 '평균 -(3×표준편차)'를, ul에는'평균+(3×표준편차)'를 저장합니다.
- Step #2) dnorm() 함수를 이용하여 $N(170, 6^2)$의 분포도를 작성합니다.

```
7. x <- seq(ll, ul, by=0.01)
8. nd <- dnorm(x, mean=mu, sd=sigma)
9. plot(x, nd, type="l", xlab="x", ylab="P(X=x)", lwd=2,
   col="red")
```

정규분포

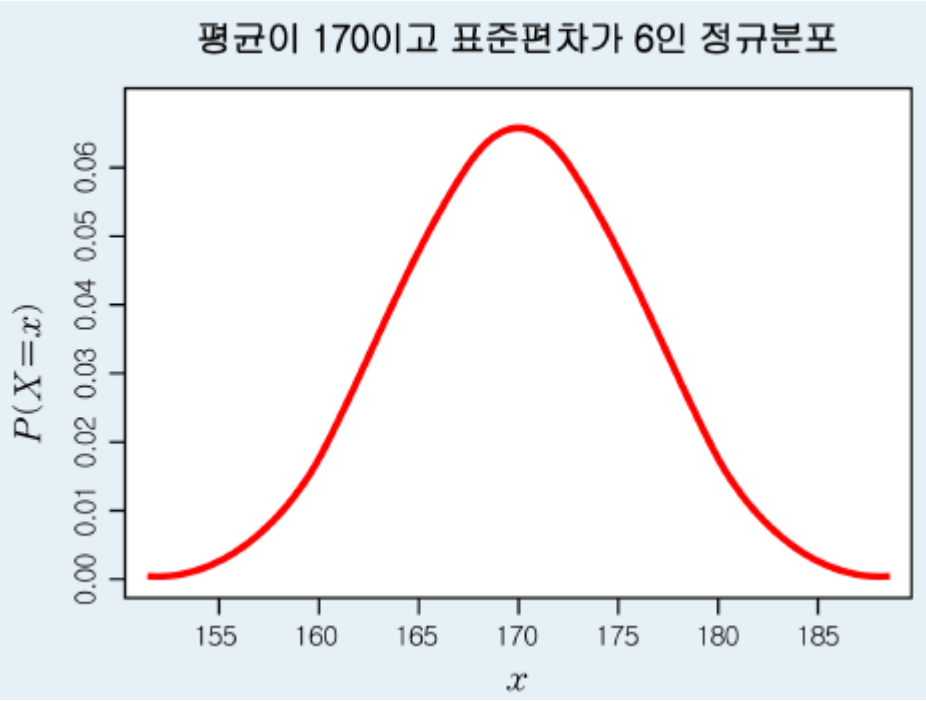
- 7줄 : 확률변수 X가 갖는 값을 ll부터 ul까지 0.01씩 증가하는 값으로 하여 벡터 x에 저장합니다.
- 8줄 : dnorm() 함수는 정규분포의 확률밀도함수를 구하는 함수입니다.

```
dnorm(x, mean, sd)
- x : 정규분포가 가질 수 있는 값의 벡터(원소 1개짜리 포함)
- mean : 정규분포의 평균
- sd : 정규분포의 표준편차
```

  - dnorm() 함수의 첫 번째 전달인자로 코드에서와 같이 벡터를 전달할 경우 각 값의 확률밀도함수 값을 갖는 벡터를 전달합니다.
  - 여기서 구한 확률밀도함수는 우리가 구하고자 하는 범위 내의 모든 x에 대한 확률($P(X = x)$)로, 이를 변수 nd 에 벡터로 저장합니다.

정규분포

- 9줄 : 위에서 구한 x를 x축의 값으로, 각 x의 확률값 nd를 높이로 하여 산점도를 그립니다.
  - 산점도의 형태를 type="l"로 하여 각값을 선으로 연결합니다.
  - 앞서 구한 x 값은 0.01씩 증가하는 이산형 자료이나, 이 증가분을 작게 하여 연속형 자료처럼 표현합니다



정규분포

- Step #3) pnorm() 함수를 이용하여 $N(170, 6^2)$ 의 분포함수를 구합니다

```
pnorm(x, mean, sd)
- x : 정규분포가 가질 수 있는 값의 벡터(원소 1개짜리 포함)
- mean : 정규분포의 평균
- sd : 정규분포의 표준편차
```

```
11.pnorm(mu, mean=mu, sd=sigma)
12.pnorm(158, mean=mu, sd=sigma)
13.pnorm(180, mean=mu, sd=sigma) -
                        pnorm(160, mean=mu, sd=sigma)
```

- 11줄 : 확률변수 X가 $N(170, 6^2)$ 을 따를 때 $P(X \leq 170)$을 구합니다.
  - 변수 mu에 평균값 170이 저장되어 있으며, 정규분포에서 평균 이하일 확률은 0.5입니다(평균을 중심으로 좌우대칭).

정규분포

- 12줄 : 확률변수 X가 $N(170, 6^2)$ 을 따를 때 $P(X \leq 158)$ 을 구합니다.
  - $N(170, 6^2)$ 에서 158은 '170-(2×6)'인 값으로, 정규분포에서 이 확률은 약 0.0228입니다.
  - 정규분포는 좌우대칭이므로 '170+(2×6)'인 182이 이상일 확률 $P(X \geq 182)$ 또한 약 0.0228입니다.
- 13줄 : 확률변수 X가 $N(170, 6^2)$ 을 따를 때 $P(X \leq 180) - P(X \leq 160)$ 즉, $P(160 \leq X \leq 180)$의 확률을 구합니다. 구한 값은 약 0.904 입니다.

```
> pnorm(mu, mean=mu, sd=sigma)
[1] 0.5
> pnorm(158, mean=mu, sd=sigma)
[1] 0.0228
> pnorm(180, mean=mu, sd=sigma) - pnorm(160, mean=mu, sd=sigma)
[1] 0.904
```

정규분포

- Step #4) qnorm() 함수를 이용하여 $N(170, 6^2)$ 의 분위 p에 해당하는 값을 구합니다

```
qnorm(p, mean, sd)
- p : 구하고자 하는 분위 벡터(원소 1개짜리 포함)
- mean : 정규분포의 평균
- sd : 정규분포의 표준편차
```

- 1
증

```
15.qnorm(0.25, mean=mu, sd=sigma)
16.qnorm(0.5, mean=mu, sd=sigma)
17.qnorm(0.75, mean=mu, sd=sigma)
```

정규분포

- Step #5) rnorm() 함수를 이용하여 $N(170, 6^2)$ 에서 400개의 난수를 생성하고 모집단과 비교해 봅니다.
  - 이 과정은 모집단이 $N(170, 6^2)$ 을 따를 때 400개의 확률표본을 추출하는 것을 R로 구현해보는 과정입니다.(이항분포에서도 마찬가지입니다.)

```
> qnorm(0.25, mean=mu, sd=sigma)
[1] 166
> qnorm(0.5, mean=mu, sd=sigma)
[1] 170
> qnorm(0.75, mean=mu, sd=sigma)
[1] 174
```

```
rnorm(n, mean, sd)
- n : 정규분포로 부터 추출할 난수의 개수
- mean : 정규분포의 평균
- sd : 정규분포의 표준편차
```

정규분포

```
19.options(digits=5)

20.set.seed(5)

21.smp <- rnorm(400, mean=mu, sd=sigma)

22.c(mean(smp), sd(smp))

23.hist(smp, prob=T, main="N(170, 6^2)으로부터 추출한 표본의 분포
   (n=400)", xlab="", ylab="", col="white", border="black")

24.lines(x, nd, lty=2, lwd=2, col="red")
```

정규분포

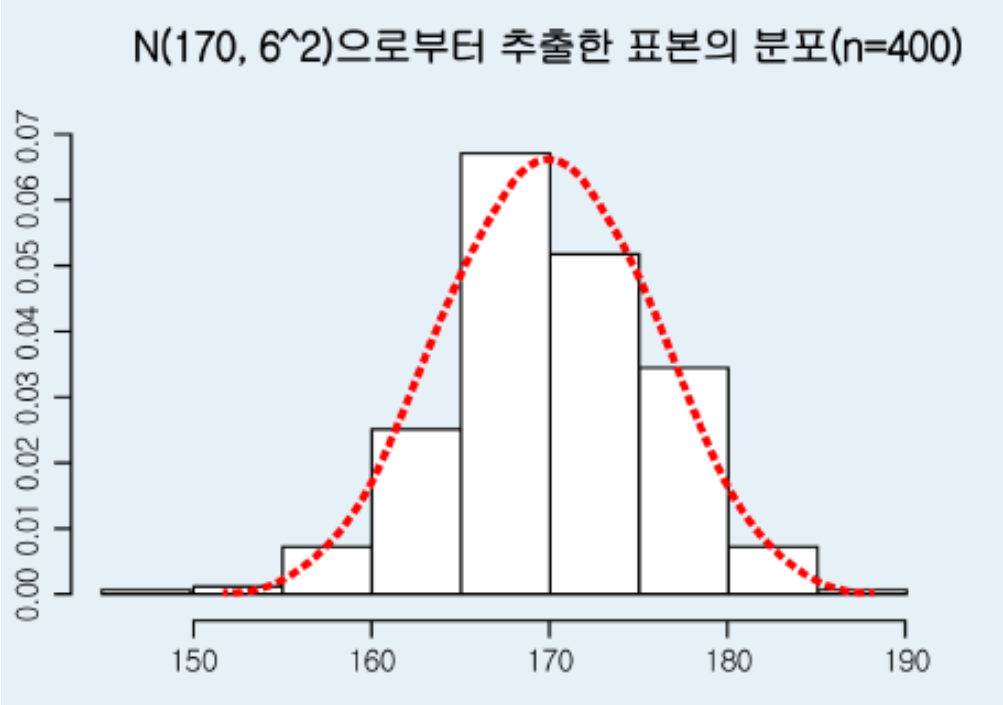
- 21줄 : 변수 smp에 $N(170, 6^2)$ 로부터 400개의 표본을 추출하고 저장합니다.
- 22줄 : 생성한 smp의 평균과 표준편차를 벡터로 출력했습니다.
  - 표본의 개수가 충분히 크다면 모집단의 평균과 표준편차와 비교해보면 크게 차이가 나지 않습니다.

```
> c(mean(smp), sd(smp))
[1] 170.0165   6.0054
```

- 23줄 : 추출한 표본의 상대도수(prob=T)로 히스토그램을 작성합니다
- 24줄 : 표본의 히스토그램 위에 점선(lty=2)으로, 선굵기는 2(lwd=2)인 붉은색(col="red")의 선으로 $N(170, 6^2)$ 의 분포를 표시했습니다.
  - 모집단의 분포와 차이는 있지만 많이 닮았음을 알 수 있습니다
  - lines() 함수와 같이 기존에 작성된 그래프 위에 각종 표현을 하는 함수를 **저수준 그래프 함수**라 합니다. lines() 함수는 전달된 좌표들을 선으로 연결합니다.
    - 첫번째와 두번째 전달인자인 x와 nd의 각 순서쌍 $(x_1, nd_1), (x_2, nd_2), \ldots$ 를 좌표로 하여 각 점들을 선으로 연결합니다.

정규분포

```
lines(x, y, …)
- x : x좌표로 사용할 벡터(원소 1개짜리 포함)
- y : y좌표로 사용할 벡터(원소 1개짜리 포함)
- … : 그래프에서 사용하는 공통 전달인자 (lty, lwd, col 등등)
```



정규분포

- 개요
  - R을 이용하여 정규분포의 특징을 알아봅시다.
    - 표준정규분포를 이용하여 하위 2.5%, 5%에 해당하는 값을 찾아봅시다.
    - 정규분포의 대칭성을 이용하여 상위 2.5%, 5%에 해당하는 값은 부호만 바뀝니다.
    - 정규분포의 분포함수를 이용하여 확률계산을 해 봅시다.
      - 그래프를 통해 어떤 면적인지 확인해 봅시다.
- Step#1) 표준정규분포의 모수 준비

```
1. options(digits = 4)
2. mu <- 0
3. sigma <- 1
```

정규분포

- 1줄 : 앞서 사용한 바와 같이 출력물이 네 자릿수가 되도록 합니다.
  - 다음의 코드를 실행 하기 전에 ex1, ex2, ex3가 어떻게 출력될지 예측하고 R에 입력하여 본인 생각과 일치하는지 확인해 봅시다.

```
ex1 <- 12.356
ex2 <- 0.001234
ex3 <- c(ex1, ex2)
```

  - 출력상 값이 digits로 정해진 방식에 따라 출력될 뿐 원래의 값은 유지하고 있습니다. 즉, options(digits=n) 으로는 값 자체를 바꾸지 않습니다.
- 2줄 : 표준정규분포의 평균인 0을 변수 mu에 저장합니다.
- 3줄 : 표준정규분포의 표준편차인 1을 변수 sigma에 저장합니다.

정규분포

- Step #2) 표준정규분포의 특별한 값을 찾습니다.

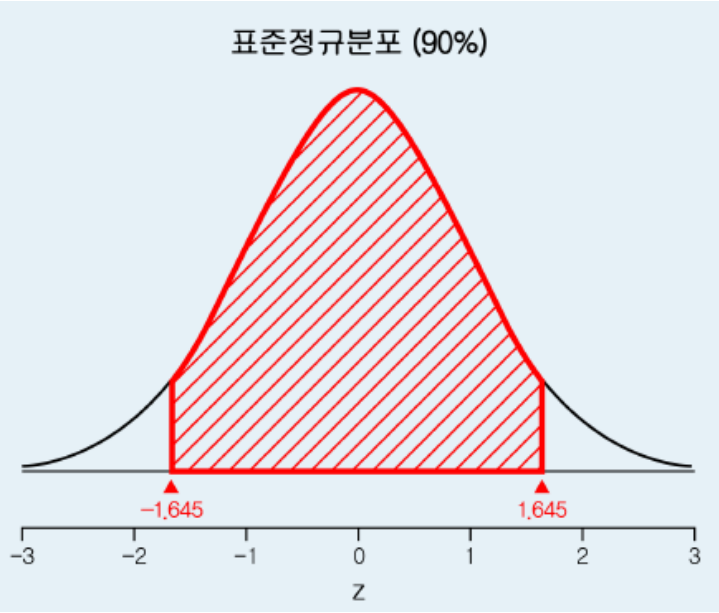
```
5. (p0.05 <- qnorm(0.05, mean=mu, sd=sigma))
6. (p0.025 <- qnorm(0.025, mean=mu, sd=sigma))
```

  - 5줄 : $P(Z \leq z) = 0.05$인 z값을 qnorm() 함수로 구합니다.
    - 표준정규분포에서 z가 -1.645보다 작을 확률은 약0.05(5%)입니다.
    - 표준정규분포는 좌우대칭으로 z가 1.645보다 클 확률 역시 약 0.05(5%)입니다.
  - 5줄 : $P(Z \leq z) = 0.025$인 z값을 qnorm() 함수로 구합니다.
    - 표준정규분포에서 z가 -1.96보다 작을 확률은 약0.025(2.5%)입니다.
    - 표준정규분포는 좌우대칭으로 z가 1.96보다 클 확률 역시 약 0.025(2.5%)입니다.

```
> (p0.05 <- qnorm(0.05, mean=mu, sd=sigma))
[1] -1.645
> (p0.025 <- qnorm(0.025, mean=mu, sd=sigma))
[1] -1.96
```

정규분포

- Step #3) 분포함수를 통해 원하는 구간의 면적을 찾습니다
  - 8 줄
    - 앞서 -1.645보다 작은 쪽의 면적은 0.05이고 1.645보다 큰 쪽의 면적 또한 0.05이므로 그 사이의 면적은 0.9임을 알 수 있습니다.
    - pnorm() 함수를 이용하여 $P(-1.645 \leq Z \leq 1.645)$를 확인합니다.
    - 정규분포에서 (평균 - 1.645×표준편차)부터 (평균 + 1.645×표준편차) 사이에 들어갈 확률은 약 90%입니다.



정규분포

- 9 줄
  - 앞서 -1.96보다 작은 쪽의 면적은 0.025이고 1.96보다 큰 쪽의 면적 또한 0.025 이므로 그 사이의 면적은 0.95임을 알 수 있습니다.
  - pnorm() 함수를 이용하여 $P(-1.96 \leq Z \leq 1.96)$를 확인해합니다.
  - 정규분포에서 (평균 - 1.96×표준편차)부터 (평균 + 1.96×표준편차) 사이에 들어갈 확률은 약 95%입니다.
- pnorm() 함수를 이용하여 다양한 정규분포의 확률을 계산할 수 있습니다.