Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

03

확률변수와 확률분포 (1)

경북대학교 배준현 교수
(joonion@knu.ac.kr)

03. 확률변수와 확률분포 (1)

- 확률변수: *random variable*
  - 확률적으로 서로 다른 값을 가질 수 있는 어떤 변수
  - 어떤 시행에서 표본공간의 각 근원사건에 하나의 실수를 대응시키는 것
    - 예) 동전 던지기의 확률변수: 동전을 던졌을 때 나오는 면
      - $X = \{H, T\}$
  - 이산확률변수: *discrete* random variable
    - 확률변수 X가 취할 수 있는 값이 불연속일 때.
  - 연속확률변수: *continuous* random variable
    - 확률변수 X가 취할 수 있는 값이 연속일 때.

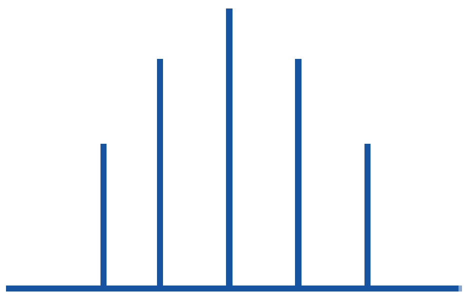
03. 확률변수와 확률분포 (1)

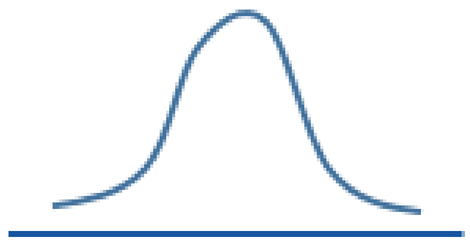
- **확률분포**: *probability distribution*
  - 확률변수 X가 갖는 값과 X가 이 값을 가질 확률 사이의 대응 관계(함수)
  - 이산확률분포: *discrete* probability distribution
    - 이산확률변수에 대한 확률질량함수: *PMF*, probability *mass* function
  - 연속확률분포: *continuous* probability distribution
    - 연속확률변수에 대한 확률밀도함수: *PDF*, probability *density* function

03. 확률변수와 확률분포 (1)

▪ 확률분포의 성질

- **이산 확률분포:**
  - $P(X = x_i) = p_i, i = 1,2,3, \dots, n$
  - $0 \le p_i \le 1$
  - $\sum_{i=1}^{n} P(X = x_i) = \sum_{i=1}^{n} p_i = 1$
- **연속 확률분포:**
  - $P(X = x) = f(x), f(x) \ge 0$
  - $P(a \le X \le b) = \int_a^b f(x)\, dx$
  - $\int_\alpha^\beta f(x)\, dx = 1, \alpha \le x \le \beta$



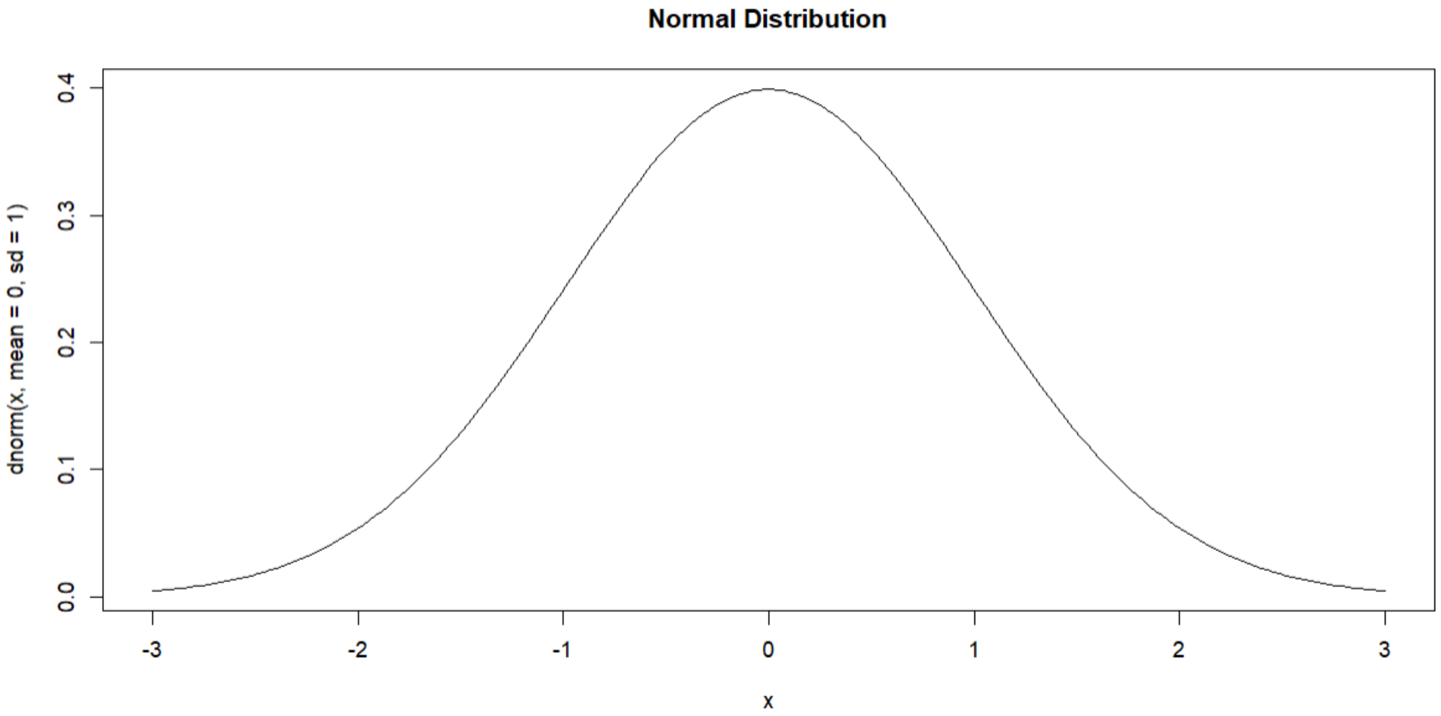


- **확률분포의 성질:**
  - **이산확률분포:**
    - **기댓값(=평균):** $E(X) = m = \sum_{i=1}^{n} x_i p_i$
    - **분산:** $(X-m)^2$ **의 평균.**
      - $V(X) = \sum_{i=1}^{n}(x_i - m)^2 p_i = \sum_{i=1}^{n} {x_i}^2 p_i - m^2$
    - **표준편차:** $\sigma(X) = \sqrt{V(X)}$
  - **연속확률분포:**
    - **기댓값(=평균):** $E(X) = \int_{\alpha}^{\beta} x f(x)\, dx$
    - **분산:** $V(X) = \int_{\alpha}^{\beta} (x-m)^2 f(x)\, dx = \int_{\alpha}^{\beta} x^2 f(x)\, dx - m^2$
    - **표준편차:** $\sigma(X) = \sqrt{V(X)}$

- **정규분포:** *normal distribution*
  - 확률밀도함수를 그래프로 그렸을 때 **종형 곡선**이 나타나는 확률분포
    - 확률밀도함수: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, -\infty < x < \infty$
    - 평균이 m, 표준편차가 σ인 정규분포: $N(m, \sigma^2)$
  - 표준정규분포: *standard* normal distribution
    - $N(0, 1^2)$: 평균이 $m = 0$, 표준편차가 $\sigma = 1$인 정규분포
    - 확률밀도함수: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, -\infty < x < \infty$

03. 확률변수와 확률분포 (1)

```r
x <- seq(-3, 3, length=200)
plot(x, dnorm(x, mean=0, sd=1), type='l', main="Normal Distribution")
```



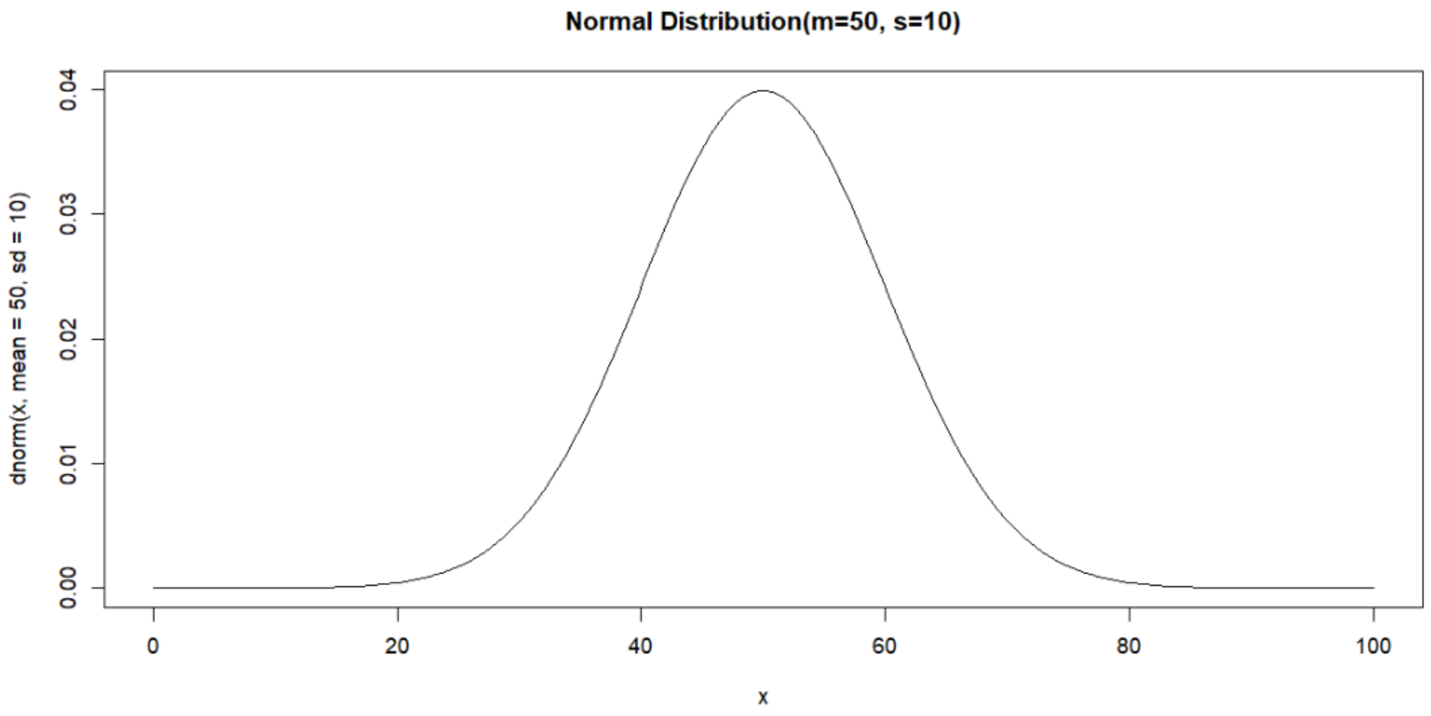
▪ 정규분포의 유용성

- 다양한 사회 현상, 자연 현상에 대한 우리의 직관과 부합하는 특성을 가짐
  - 대부분의 데이터는 평균을 중심으로 가까이 모여 있거나
  - 평균에서 양이나 음의 방향으로 떨어진 정도가 대개 비슷하거나
  - 평균에서 많이 떨어진 값들은 그리 많이 존재하지 않는다.
- 모든 (확률) 모형은 틀렸다. 하지만 그 중 어떤 것은 유용하다.
  - *All models are wrong, but some are useful*. feat. by George Box.
  - 확률분포는 어디까지나 이론적 단순화에 불과하고
  - 현실에서 발견되는 데이터가 특정 확률분포에 완전히 부합하지 않는다.

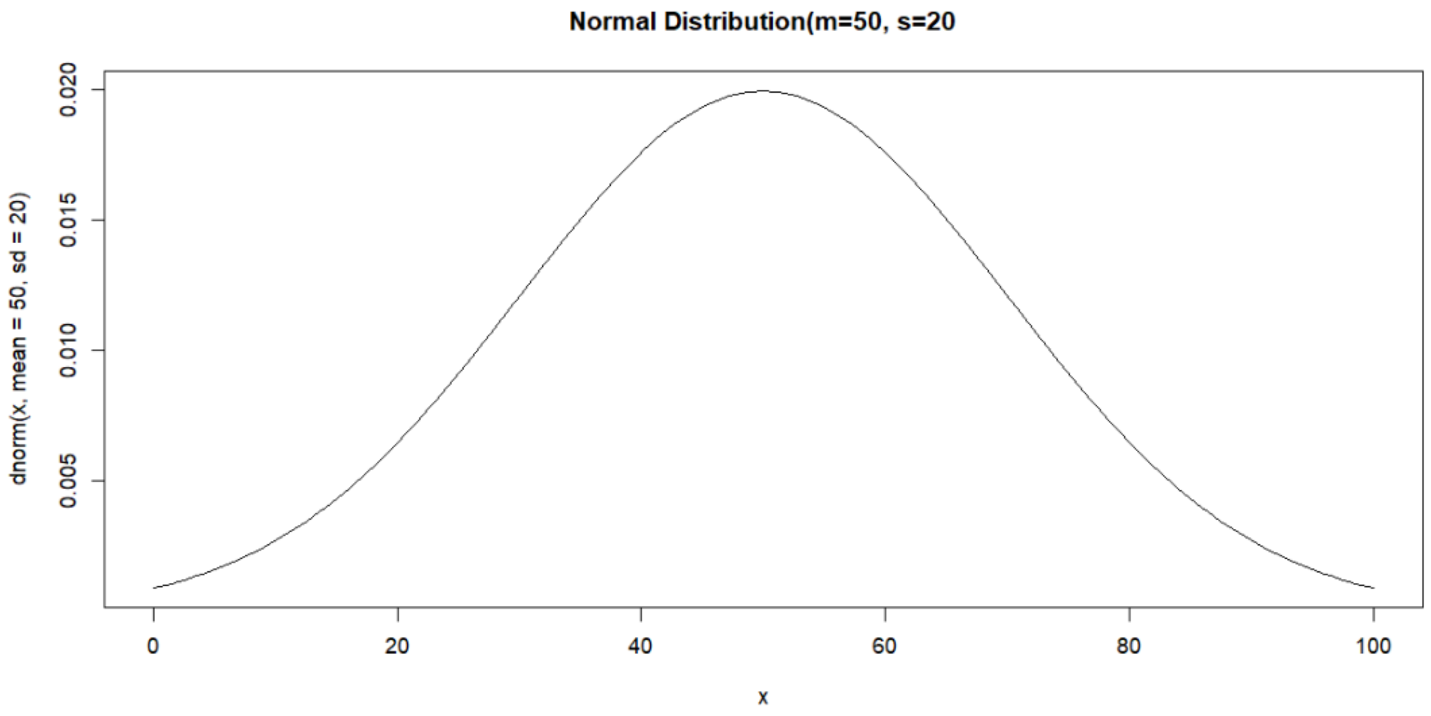
▪ 정규분포의 특성

- 기댓값(평균): 데이터의 분포를 숫자 하나로 요약할 수 있음
  - 정규분포곡선은 봉우리가 하나밖에 없고, 대부분의 값이 평균 주변에 있음
  - 정규분포곡선은 평균을 중심으로 대칭 구조임
- 표준편차: 평균에서 자료가 얼마나 떨어져 있는지를 나타내는 값
  - 표준편차 안에 들어오는 값의 비율이 항상 일정함
    - $1 \times$ 표준편차 안에 들어오는 값의 비율: *70%*
    - $2 \times$ 표준편차 안에 들어오는 값의 비율: *95%*
  - 표준편차는 정규분포곡선의 모양과 관련이 있음
    - 표준편차가 클수록 그래프는 낮고 납작해지며
    - 표준편차가 작을수록 그래프는 높고 뾰족해짐

```
x <- seq(0, 100, length=200)
plot(x, dnorm(x, mean=50, sd=10), type = 'l', main="Normal Distribution(m=50, s=10)")
```



```r
x <- seq(0, 100, length=200)
plot(x, dnorm(x, mean=50, sd=20), type = 'l', main="Normal Distribution(m=50, s=20")
```

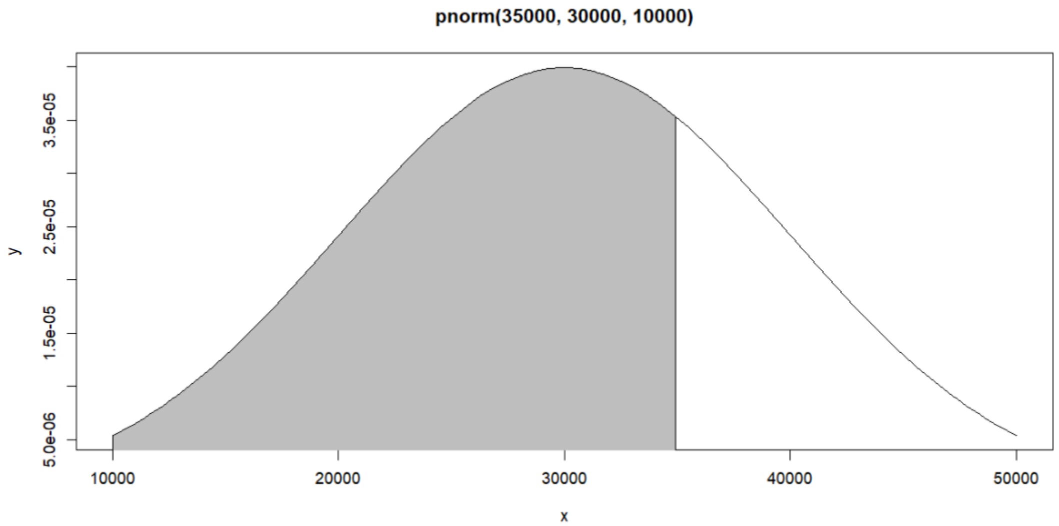


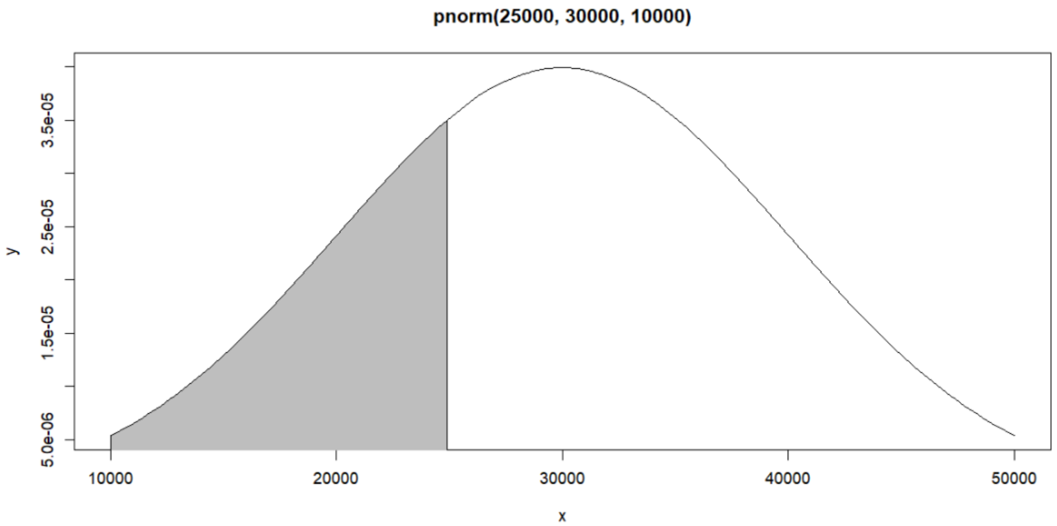
연습문제:

- 국민소득이 평균이 \$30,000, 표준편차가 \$10,000인 정규분포를 따른다고 가정한다. 즉, X를 개인의 소득을 나타내는 확률변수라 할 때,
  - $X \sim N(30000, 10000^2)$
- 어떤 사람의 소득이 \$25,000 ~ \$35,000 사이에 있을 확률을 구하시오.

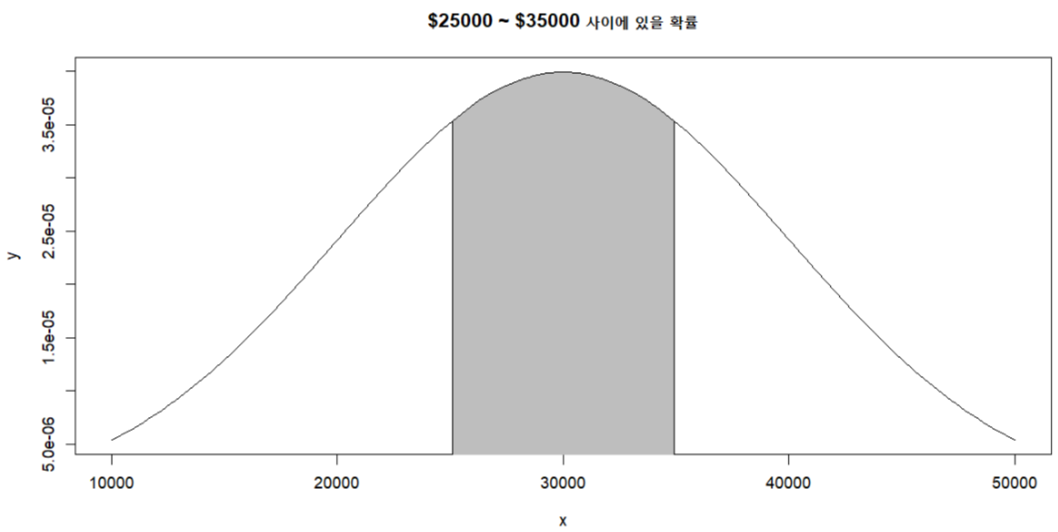
- `pnorm(q, mean=0, sd=1)`
  - `q`: vector of quantiles
  - `mean`: vector of means
  - `sd`: vector of standard deviations

```
pnorm(35000, 30000, 10000) - pnorm(25000, 30000, 10000)
```



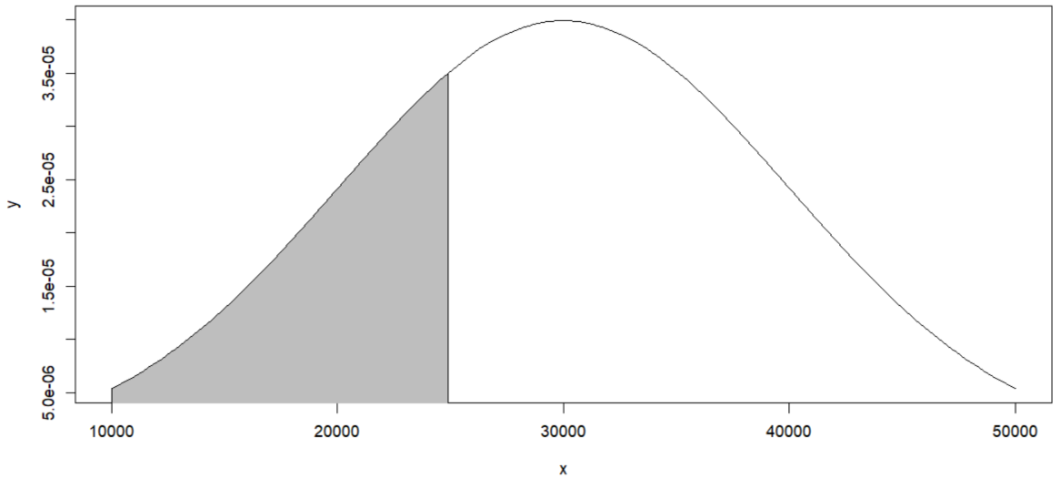


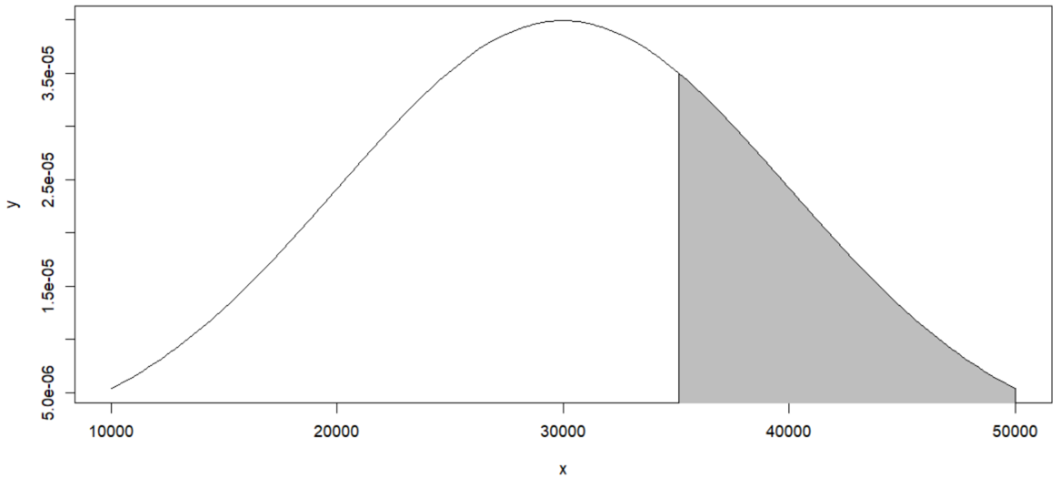
03. 확률변수와 확률분포 (1)




```r
x <- seq(10000, 50000, length=200)
y <- dnorm(x, mean = 30000, sd = 10000)
plot(x, y, type='l',
     main="$25000 ~ $35000 사이에 있을 확률")
xlim <- x[25000<=x & x<=35000]
ylim <- y[25000<=x & x<=35000]
xlim <- c(xlim[1], xlim, tail(xlim, 1))
ylim <- c(0, ylim, 0)
polygon(xlim, ylim, col="grey")
```

- **연습문제** 15.1:
  - **소득이** $25,000**보다 작을 확률은**?
    - `pnorm(25000, 30000, 10000)=0.3085375`
  - **소득이** $35,000**보다 클 확률은**?
    - `1-pnorm(35000, 30000, 10000)=0.3085375`





03. 확률변수와 확률분포 (1)

- **표준화(정규화):** *standardization*
  - 정규분포를 따르는 확률변수 X를 표준정규분포를 따르는 확률변수 Z로 변환
    - $X \sim N(m, \sigma^2) \rightarrow Z = \frac{X-m}{\sigma}, Z \sim N(0, 1^2)$
  - 표준화를 하는 이유:
    - 평균과 표준편차가 다른 정규분포를 따르는 두 변수의 값을 비교하는 경우

▪ 연습문제:

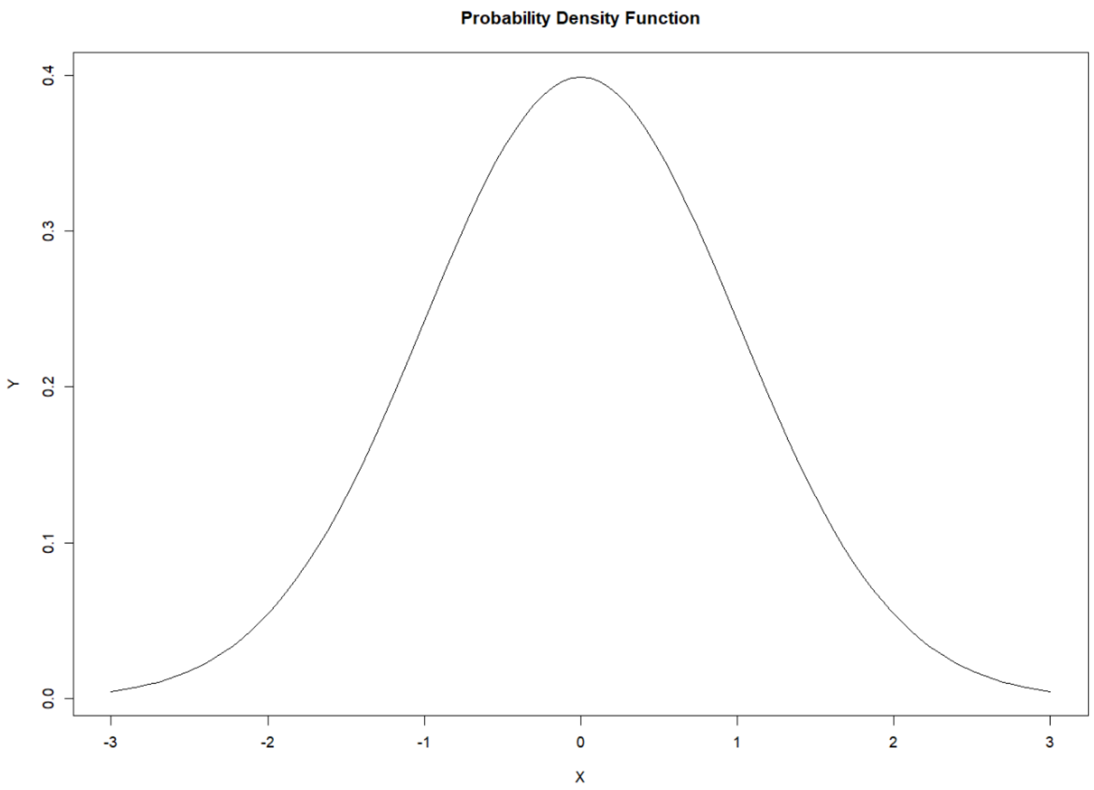
- 수학이 70점이고 영어가 80점인 학생은 어느 과목을 더 잘할까?
  - 단, 수학 점수$\sim N(60, 10^2)$, 영어 점수$\sim N(70, 20^2)$

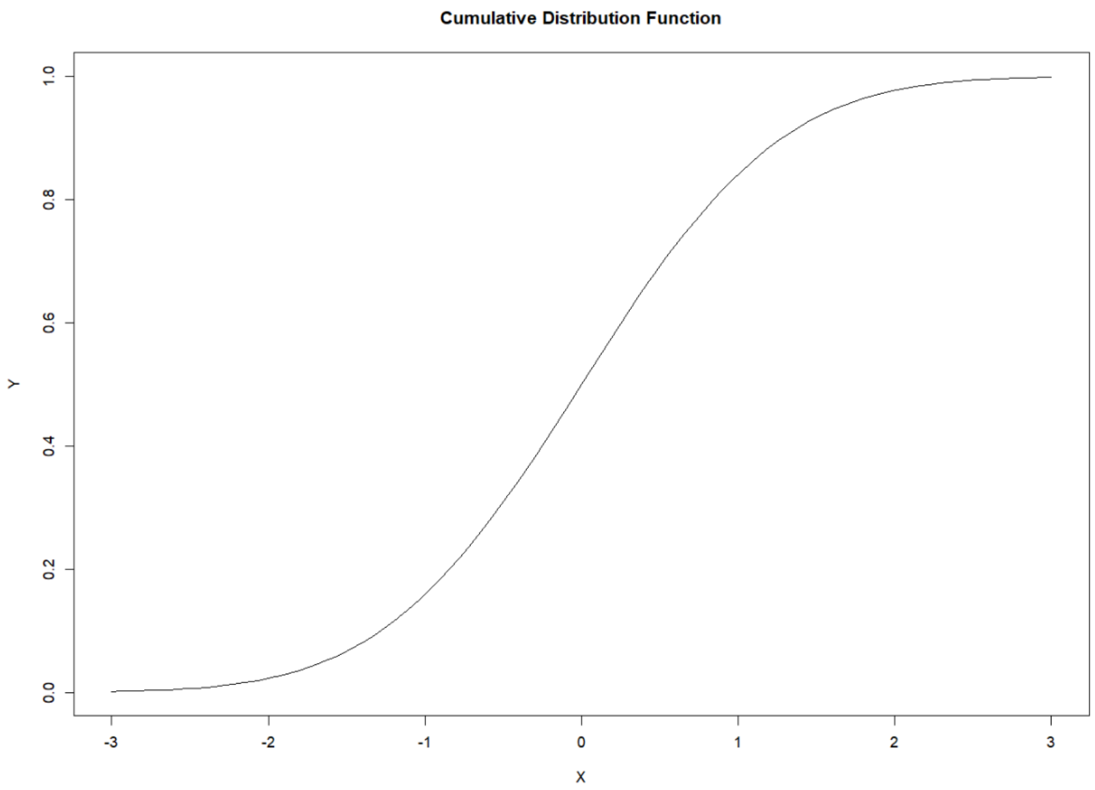
```
1 - pnorm(70, 60, 10)
1 - pnorm(80, 70, 20)

z1 <- (70 - 60) / 10
z2 <- (80 - 70) / 20
z1
z2

1 - pnorm(z1)
1 - pnorm(z2)
```

03. 확률변수와 확률분포 (1)

```r
curve(dnorm(x), -3, 3,
      xlab = 'X', ylab = 'Y',
      main="Probability Density Function")
```

```r
curve(pnorm(x), -3, 3,
      xlab = 'X', ylab = 'Y',
      main="Cumulative Distribution Function")
```





▪ 이항 분포: *binomial distribution*

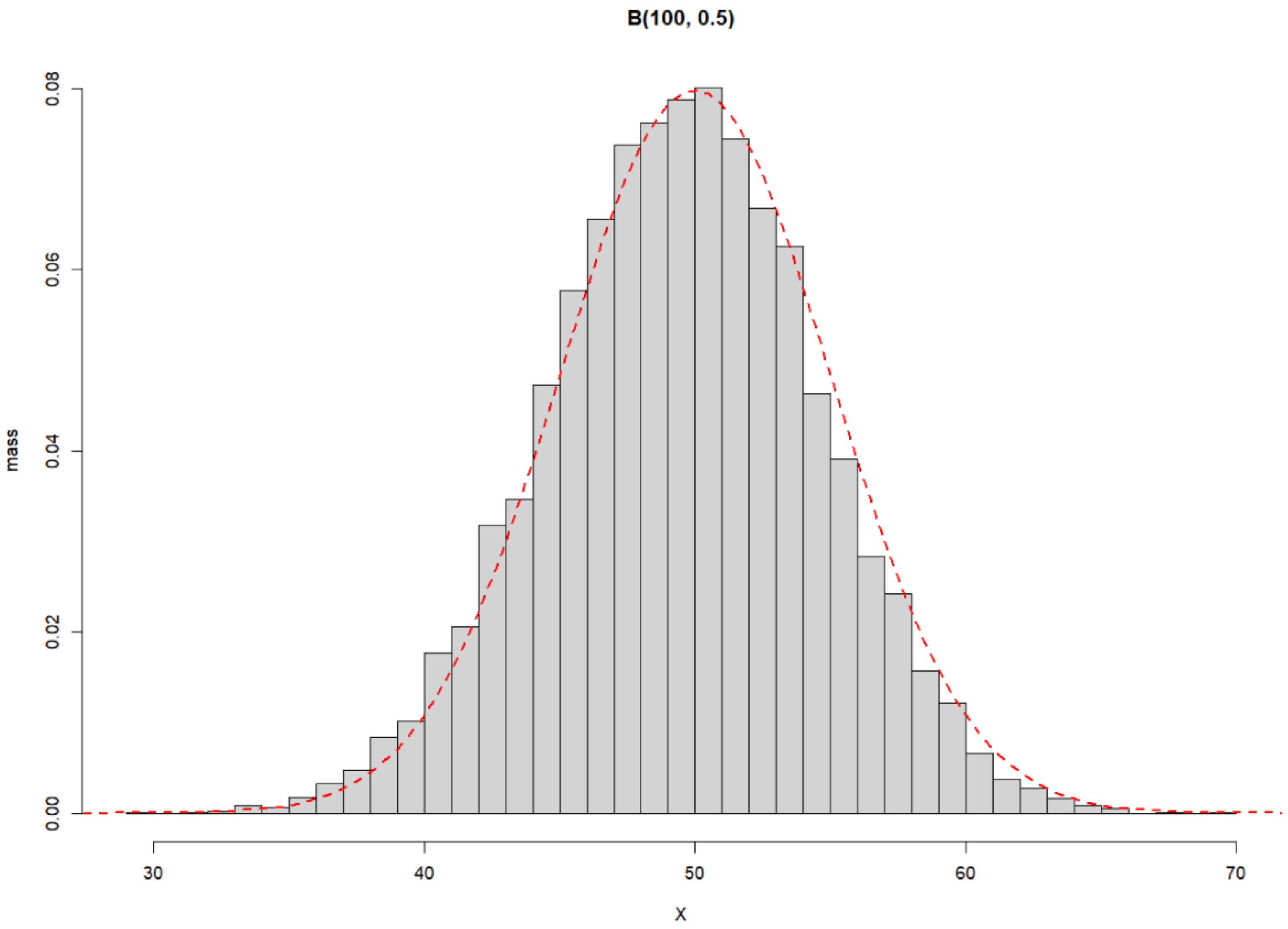
- 어떤 시행에서 사건이 일어날 확률이 p인 독립시행을 n회 반복할 때
  - 사건이 일어나는 횟수인 확률변수 X는 이항분포 $B(n, p)$를 따른다.
  - $X \sim B(n, p)$
- X의 확률질량함수
  - $P(X = r) = {}_nC_r\, p^r (1-p)^{n-r},\ r = 0,1,2,\cdots,n$

- **연습문제:**
  - 동전의 앞면이 나올 확률이 0.5일 때 동전 던지기를 100회 시행했다.
    - 동전이 앞면이 나오는 횟수를 X라고 할 때 확률분포의 그래프를 그려보자.

  - 앞면이 0번 나올 확률: $P(X=0) = {}_{100}C_0\, 0.5^0(1-0.5)^{100} = 0.5^{100}$
  - 앞면이 1번 나올 확률: $P(X=1) = {}_{100}C_1\, 0.5^1(1-0.5)^{99} = 100 \times 0.5^{100}$
  - 앞면이 2번 나올 확률: $P(X=2) = {}_{100}C_2\, 0.5^2(1-0.5)^{98} = 100 \times 99 \times 0.5^{100}$
  - ……

```r
n_sim <- 10000
y <- rbinom(n_sim, 100, 0.5)

hist(y, xlab='X', ylab='mass',
     main = 'B(100, 0.5)',
     prob = T,
     breaks = 30)

curve(dnorm(x, 50, 5), 25, 75,
      lty=2, lwd=2, col='red',
      add=T)
```

03. 확률변수와 확률분포 (1)



Any Questions?