Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

04

확률변수와 확률분포 (2)

경북대학교 배준현 교수
(joonion@knu.ac.kr)

04. 확률변수와 확률분포 (2)

- **확률변수와 확률분포:**
  - 확률변수: *random* variable
    - 확률 사건에서 나타날 수 있는 개개의 결과를 수로 나타낸 것(X로 표기)
    - 이산 확률변수: *discrete* random variable
    - 연속 확률변수: *continuous* random variable
  - 확률분포: *probability distribution*
    - 확률변수 X가 취하는 값에 대응하는 확률을 나타내는 함수
    - 확률 질량함수: *probability mass function*, PMF
    - 확률 밀도함수: *probability density function*, PDF

04. 확률변수와 확률분포 (2)

- **확률분포 관련 R 함수:**
  - *d*: *density*, *p*: *probability*, *q*: *quantile*, *r*: *random*

구분	균일분포	이항분포	정규분포	t-분포	F-분포	χ^2-분포
난수생성함수	runif()	rbinom()	rnorm()	rt()	rf()	rchisq()
확률밀도함수	dunif()	dbinom()	dnorm()	dt()	df()	dchisq()
누적확률함수	punif()	pbinom()	pnorm()	pt()	pf()	pchisq()
백분위수함수	qunif()	qbinom()	qnorm()	qt()	qf()	qchisq()

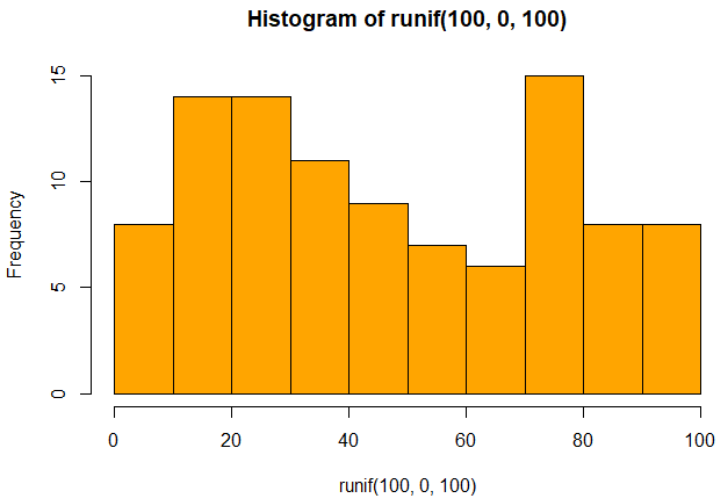
- **균일분포**: *uniform* distribution
  - **특정 범위 내에서 균등하게 나타나는 확률을 가지는 확률분포**
  - `runif(n, min, max)`
    - `n`: number of observations.
    - `min`: *lower limits* of the distribution.
    - `max`: *upper limits* of the distribution.

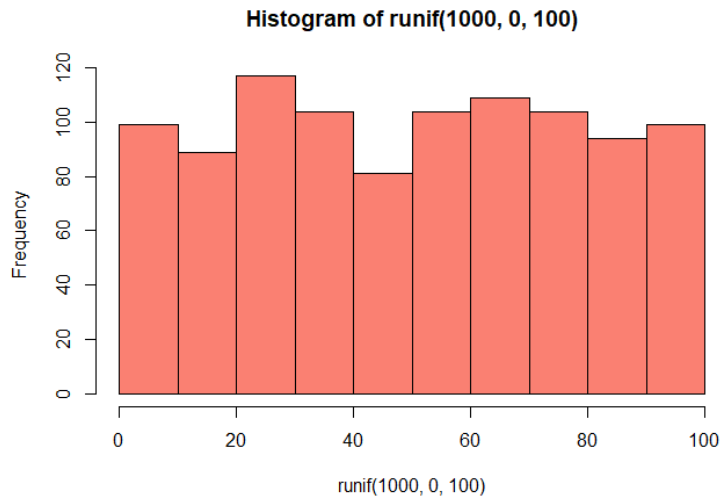
04. 확률변수와 확률분포 (2)

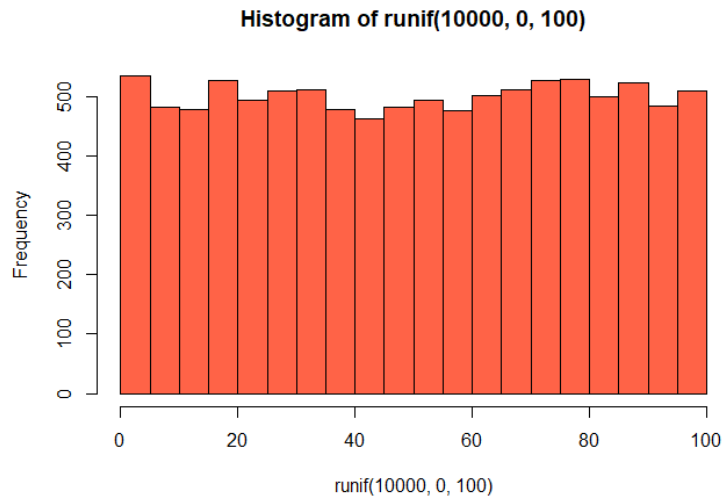
```
> runif(n = 100, min = 0, max = 100)
  [1] 95.479035 20.320243 84.033533 81.118680 57.407251 20.737249 89.220451
  [8]  8.458627 62.148120 98.383308 73.751986 39.893363 50.924797 18.181921
 [15] 88.057451 56.186187 32.507760 62.124776 66.126849 82.864863  9.037614
 [22] 38.116064  9.818295  1.122105 37.863253 84.530111 56.218531 49.242266
 [29] 20.216395 82.892048 49.020723 55.360767 12.769568 50.889883 85.815217
 [36]  5.200501 98.967980 94.456884 66.384255 89.630815 44.553949 70.167413
 [43] 90.797643  1.846722 66.877593 64.399591 52.648166 96.216954 90.399564
 [50] 42.374587 41.712281 90.293818 84.572114 91.239919 91.403600 91.181807
 [57] 93.012141  5.159144 81.254305 73.052759 15.483638 84.054914 96.931838
 [64] 18.853773 85.908074 66.309818 14.132001 98.466732  1.406115 11.054975
 [71] 70.351759 35.050807 91.048270 16.456767 91.697431 21.654728 99.699217
 [78] 13.575526 48.167717  8.461992 77.128499 32.622256 60.273190 48.896669
 [85] 66.577706 52.738507 35.634015 62.075320 99.805795 59.744214 27.190135
 [92] 69.774824  3.010868 76.182225 76.373126 94.326866 59.925745 34.002807
 [99] 67.975827 80.336392
```

04. 확률변수와 확률분포 (2)

```
> hist(runif(100,   0, 100), col = "orange")
> hist(runif(1000,  0, 100), col = "salmon")
> hist(runif(10000, 0, 100), col = "tomato")
```

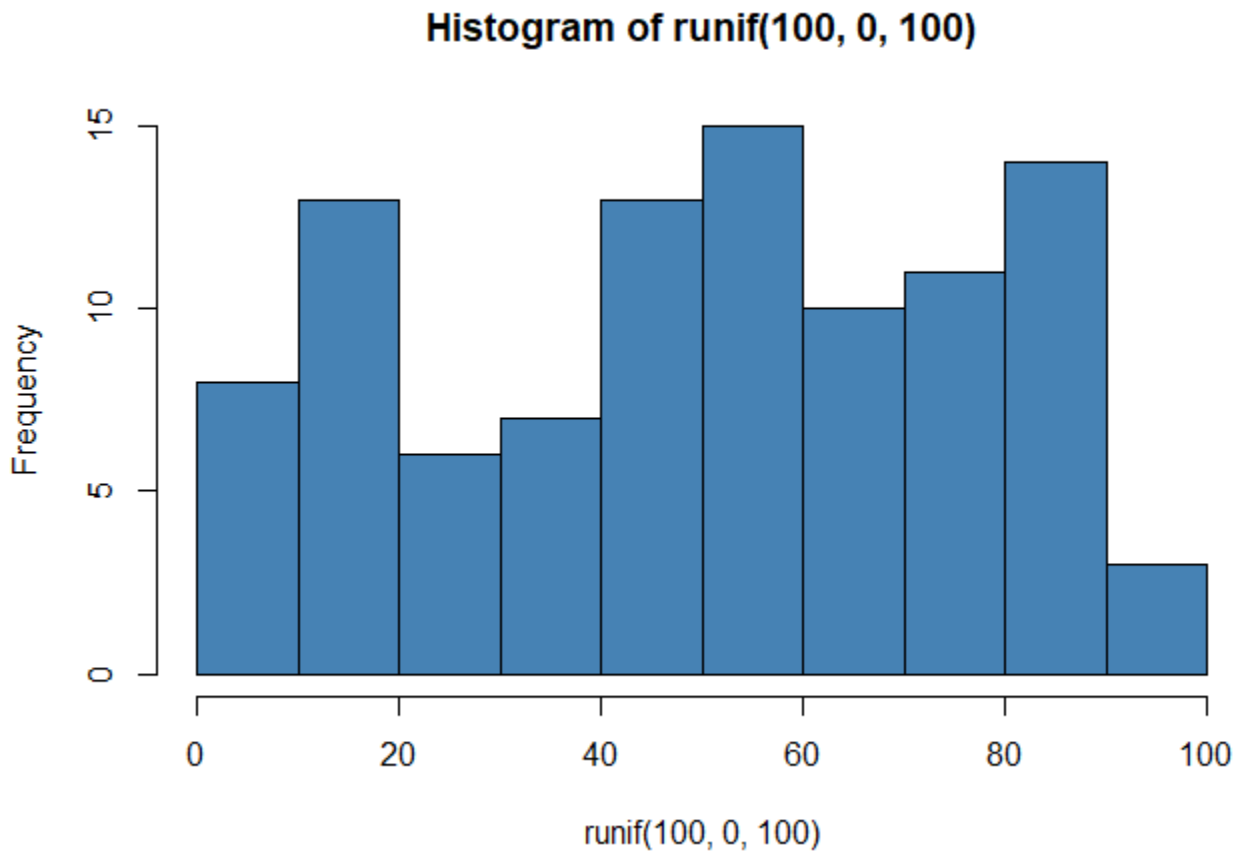






04. 확률변수와 확률분포 (2)

```
> set.seed(2022)
> hist(runif(100, 0, 100), col = "steelblue")
```



- **이항분포**: *binomial distribution*
  - 베르누이 시행: *Bernoulli trial*
    - 임의의 시행 결과가 성공 또는 실패 중 하나인 시행
  - 이항분포: *binomial distribution*
    - 성공확률이 p인 베르누이 시행을 독립적으로 n번 반복하여 시행했을 때
    - 시행의 결과가 성공인 시행의 횟수 X에 대한 확률분포
    - $X \sim B(n,p)$: 확률변수 X가 이항분포 $B(n,p)$를 따름

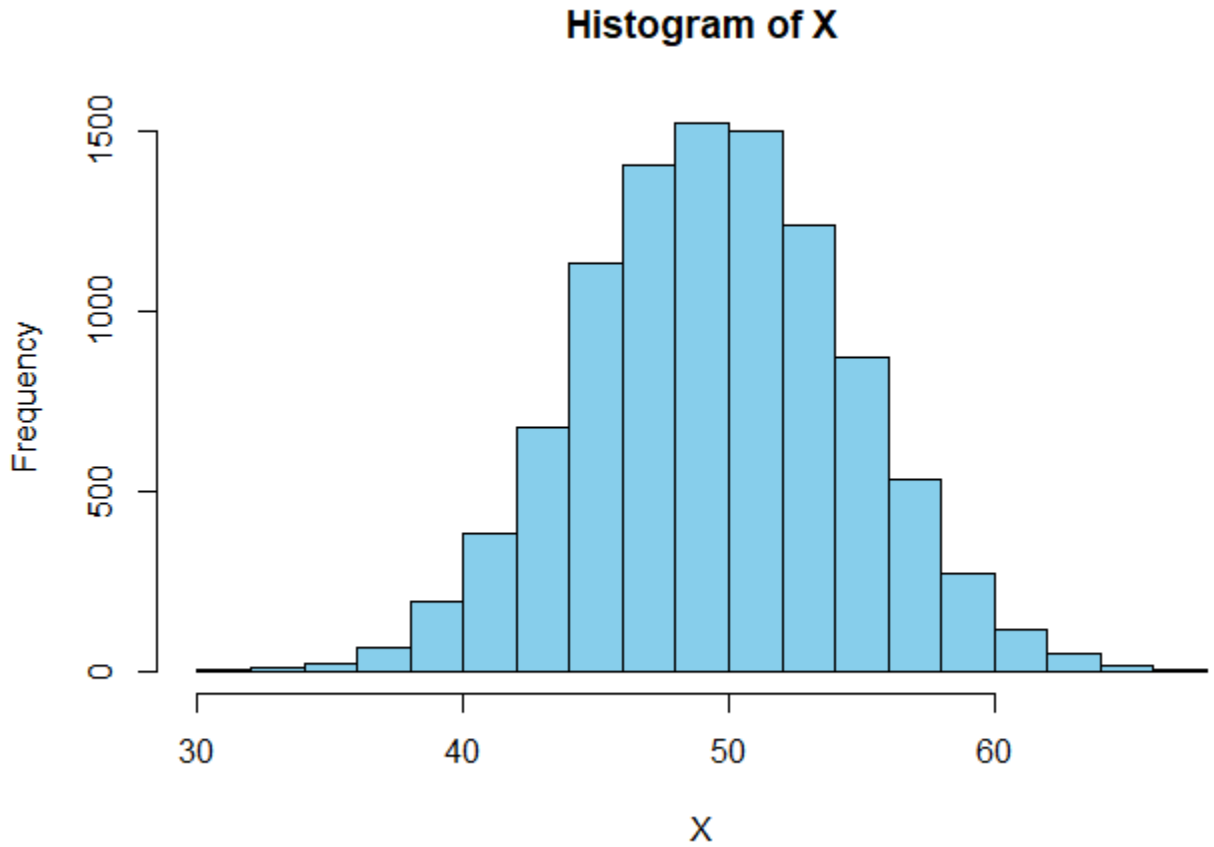
04. 확률변수와 확률분포 (2)

- X: 공평한 동전 던지기를 $size$번 실행했을 때 앞면이 나온 횟수 (성공=앞면, 성공확률 $p = 0.5$)

```
> set.seed(2022)
> rbinom(n = 1, size = 1, prob = 0.5)
[1] 1

> rbinom(n = 1, size = 10, prob = 0.5)
[1] 6

> rbinom(n = 10, size = 10, prob = 0.5)
[1] 3 5 4 6 3 2 4 6 1 3

> rbinom(n = 100, size = 10, prob = 0.5)
  [1] 3 5 5 3 6 5 5 7 7 5 5 3 3 7 5 9 5 6 5 6 4 5 3 3 7 5 6 6 6 5 6 5 7 3 7 6
 [37] 4 7 4 4 6 4 6 5 4 5 5 3 5 5 6 6 3 7 5 5 4 5 7 7 7 6 5 3 5 6 5 5 5 5 6 6
 [73] 3 5 3 4 5 6 4 3 6 5 5 4 3 9 6 6 8 4 7 6 5 4 7 3 5 8 6 5
```

04. 확률변수와 확률분포 (2)

```
> set.seed(2022)
> X <- rbinom(n = 10000, size = 100, prob = 0.5)
> hist(X, col = "skyblue", breaks = 15)
```

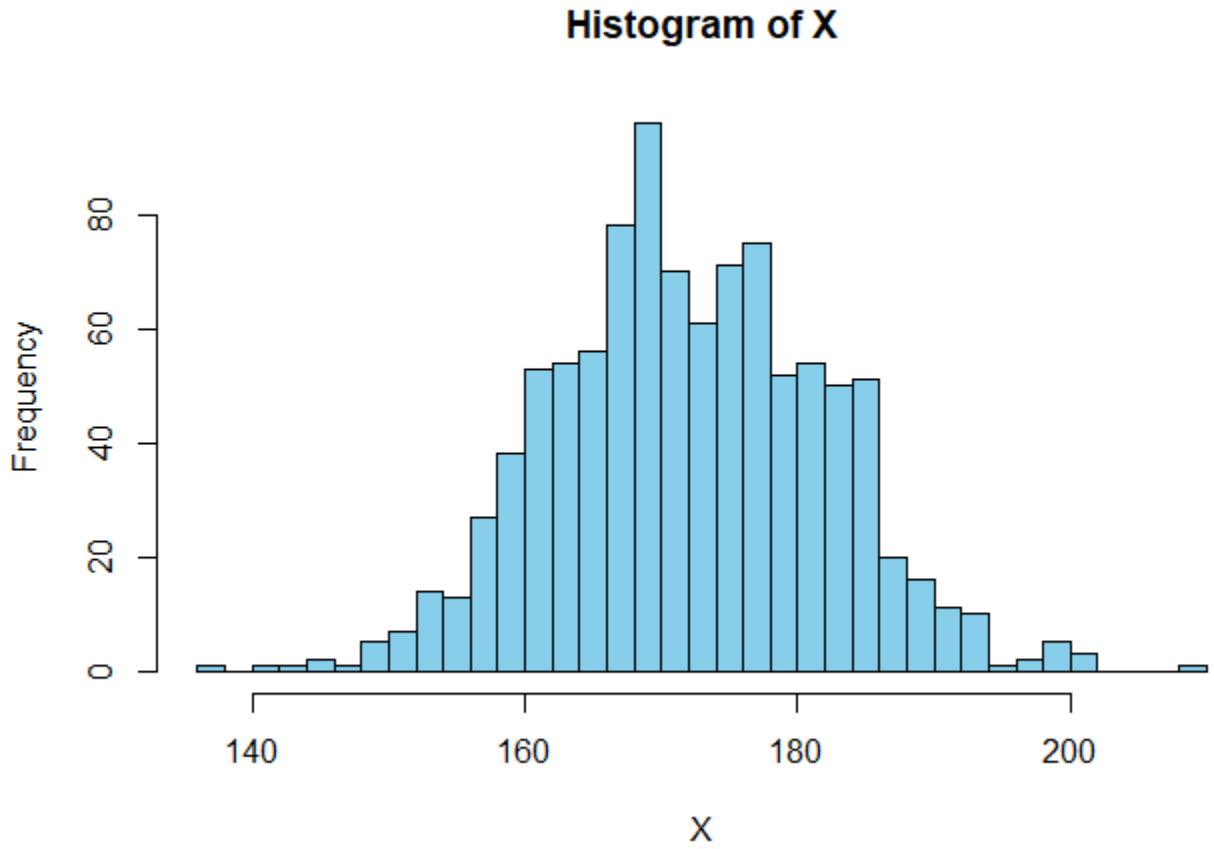


- **정규분포**: *normal distribution*
  - 자연현상 또는 사회현상에서 자주 관찰되는 **종 모양**의 확률분포
    - 정규분포는 평균과 분산(표준편차)에 따라 분포의 형태가 결정됨
    - $X \sim N(\mu, \sigma^2)$: X는 **평균**이 μ이고, **표준편차**가 σ인 정규분포를 따름
  - `rnorm(n, mean, sd)`
    - `n`: number of observations.
    - `mean`: vector of *means*.
    - `sd`: vector of *standard deviations*.

04. 확률변수와 확률분포 (2)

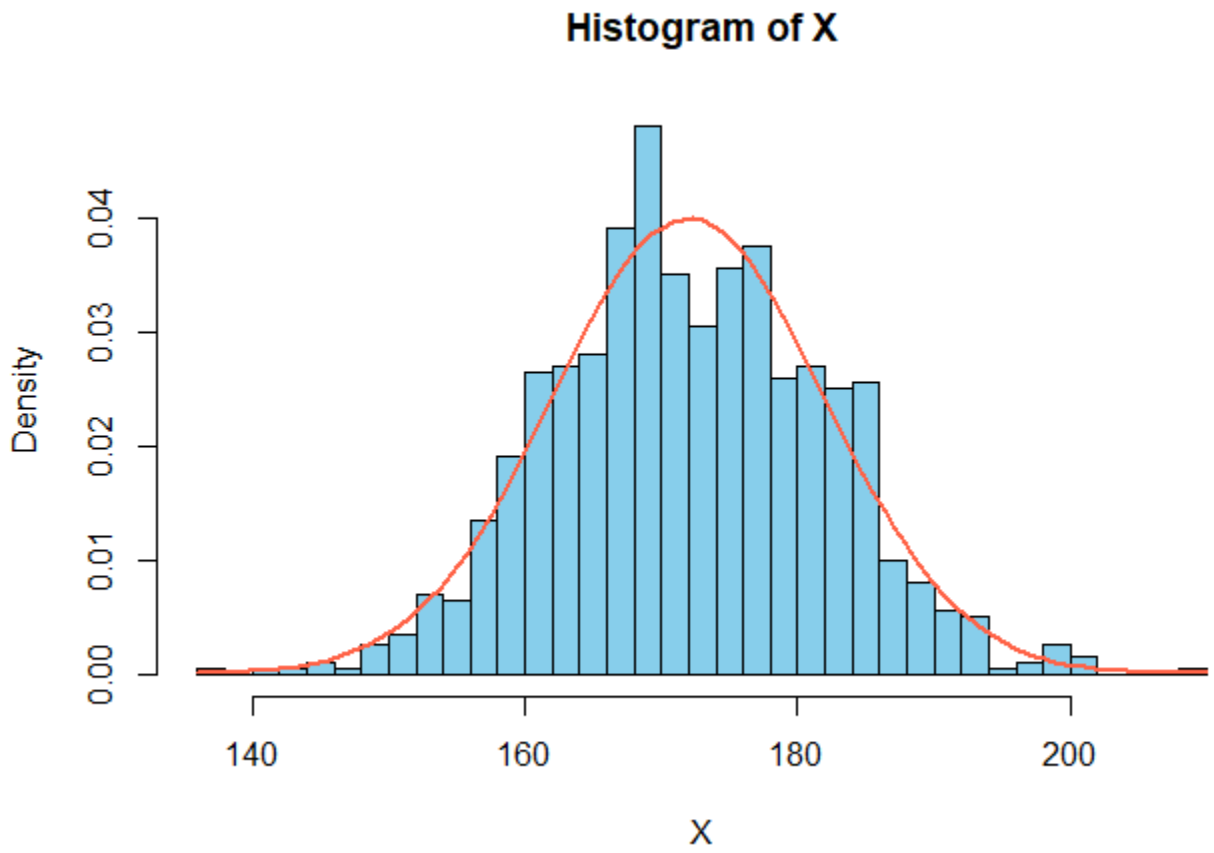
- $X \sim N(172, 10^2)$: **경북대 대학원 학생들의 키 (평균이 172, 표준편차가 10이라고 알려진 경우)**

```
> set.seed(2022)
> X <- rnorm(n = 1000, mean = 172, sd = 10)
> hist(X, col = "skyblue", breaks = 30)
```



04. 확률변수와 확률분포 (2)

```
> hist(X, col = "skyblue", breaks = 30, freq = F)
> x <- seq(min(X), max(X), length.out = 200)
> curve(dnorm(x, 172, 10), add = T, col = "tomato", lwd = 2)
```



04. 확률변수와 확률분포 (2)

- $X \sim N(172, 10^2)$일 때 어떤 대학원 학생의 키가 160보다 크거나 180보다 작을 확률은?

```
> pnorm(q = 160, mean = 172, sd = 10)
[1] 0.1150697
> pnorm(q = 160, mean = 172, sd = 10, lower.tail = F)
[1] 0.8849303
> pnorm(q = 180, mean = 172, sd = 10)
[1] 0.7881446
> pnorm(q = 180, mean = 172, sd = 10, lower.tail = F)
[1] 0.2118554

> 1 - pnorm(160, 172, 10) - pnorm(180, 172, 10, lower.tail = F)
[1] 0.6730749

> 1 - pnorm(162, 172, 10) - pnorm(182, 172, 10, lower.tail = F)
[1] 0.6826895
> 1 - pnorm(152, 172, 10) - pnorm(192, 172, 10, lower.tail = F)
[1] 0.9544997
```

04. 확률변수와 확률분포 (2)

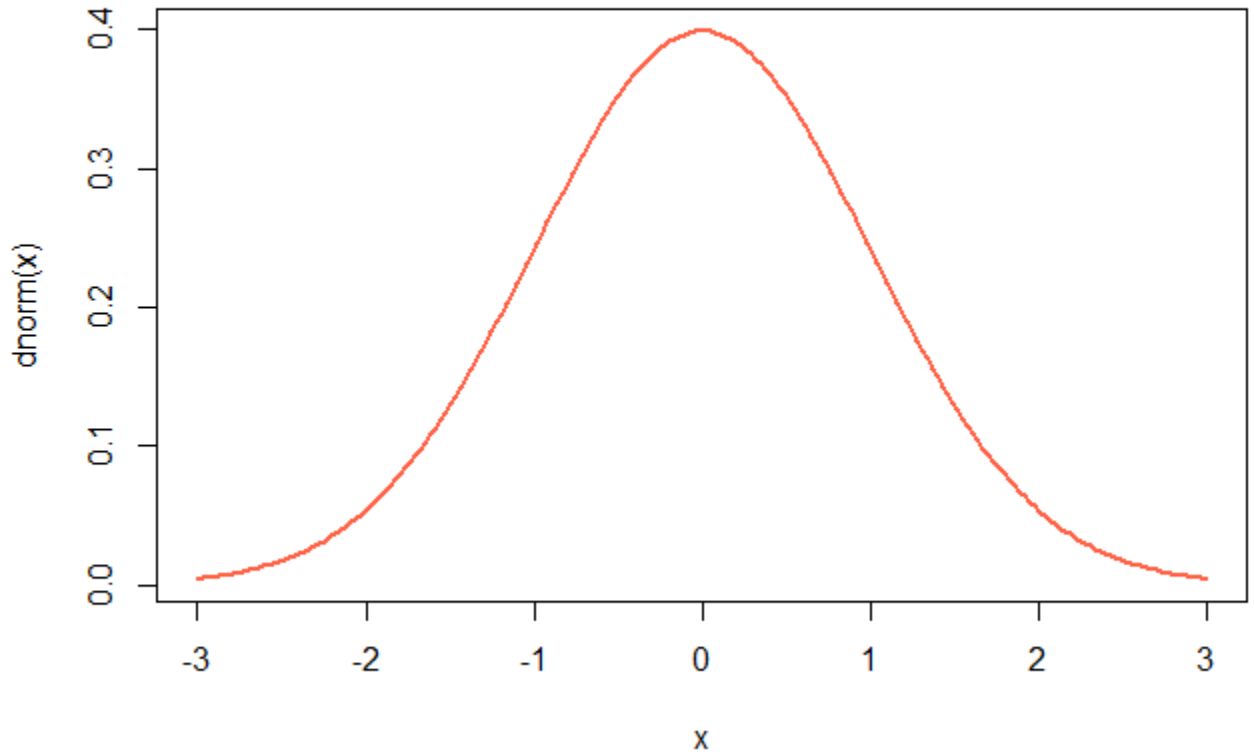
- $X \sim N(172, 10^2)$일 때 상위 5% 또는 하위 5%에 속하는 대학원생의 키는?

```
> qnorm(p = 0.05, mean = 172, sd = 10)
[1] 155.5515

> qnorm(p = 0.95, mean = 172, sd = 10)
[1] 188.4485

> qnorm(c(0.05, 0.95), 172, 10)
[1] 155.5515 188.4485

> qnorm(c(0.025, 0.975), 172, 10)
[1] 152.4004 191.5996

> qnorm(c(0.005, 0.995), 172, 10)
[1] 146.2417 197.7583
```

04. 확률변수와 확률분포 (2)

- **표준 정규분포**: $Z \sim N(0,1)$: **평균이 0이고 표준편차가 1인 정규분포 (확률변수를 Z로 표기)**

```
x <- seq(from = -3, to = 3, length.out = 200)
plot(x, dnorm(x), type = "l", col = "tomato", lwd = 2)
```

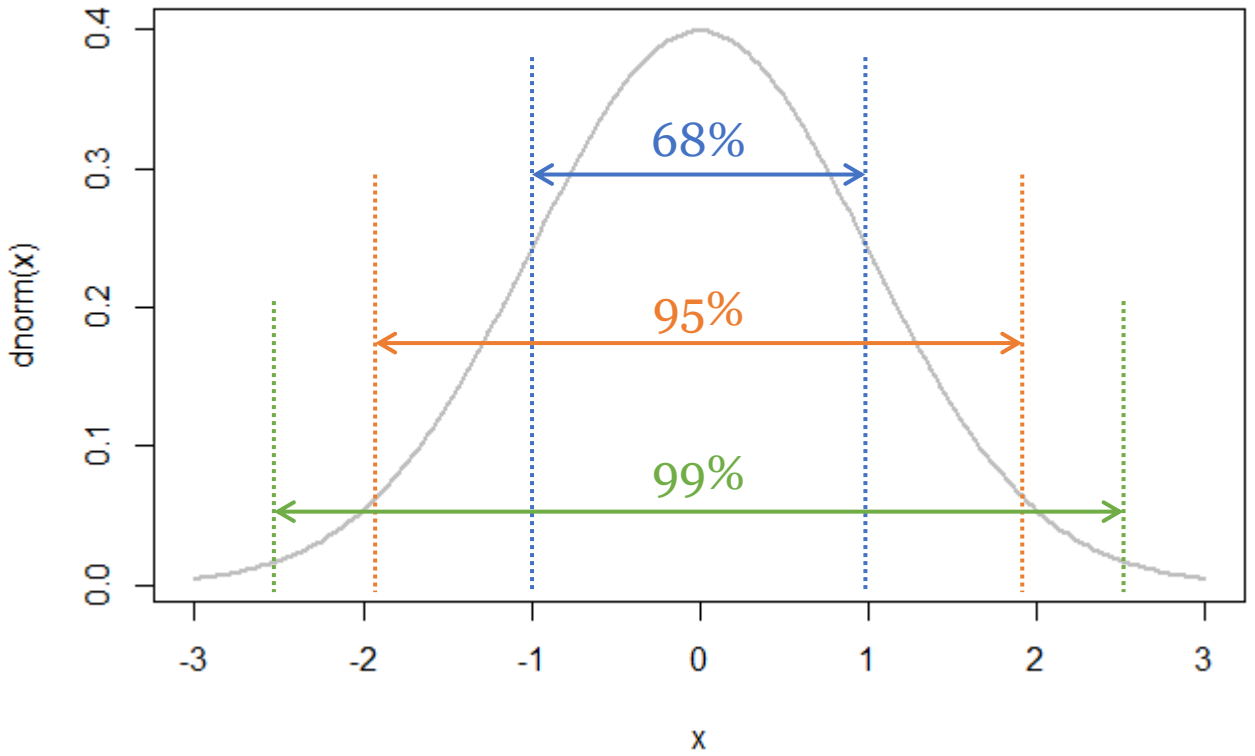


04. 확률변수와 확률분포 (2)

- 정규분포를 따르는 확률변수의 값은 $\mu \pm 1.96 \times \sigma$ 범위 안에 있을 확률이 약 95%이다.

```
> qnorm(c(0.025, 0.975), 0, 1)
[1] -1.959964  1.959964
> pnorm(c(-1.96, 1.96), 0, 1)
[1] 0.0249979 0.9750021

> qnorm(c(0.005, 0.995), 0, 1)
[1] -2.575829  2.575829
> pnorm(c(-2.58, 2.58), 0, 1)
[1] 0.004940016 0.995059984

> 1 - pnorm(-1) - pnorm(1, lower.tail = F)
[1] 0.6826895
> 1 - pnorm(-1.96) - pnorm(1.96, lower.tail = F)
[1] 0.9500042
> 1 - pnorm(-2.58) - pnorm(2.58, lower.tail = F)
[1] 0.99012
```

04. 확률변수와 확률분포 (2)



- **모집단과 표본집단: *population* and *samples***
  - 모집단: 연구의 대상이 되는 **전체 집합**
    - **모집단 분포**: 모집단의 데이터가 가지는 확률분포
  - 표본집단: 모집단으로부터 추출한 **부분 집합**
    - **표본분포**: 모집단에서 추출한 표본 데이터가 가지는 확률분포
  - 표본추출: *sampling*
    - **복원추출**: 추출한 표본을 되돌려 놓고 다음 표본을 추출
    - **비복원추출**: 이미 추출한 표본은 제외하고 다음 표본을 추출

04. 확률변수와 확률분포 (2)

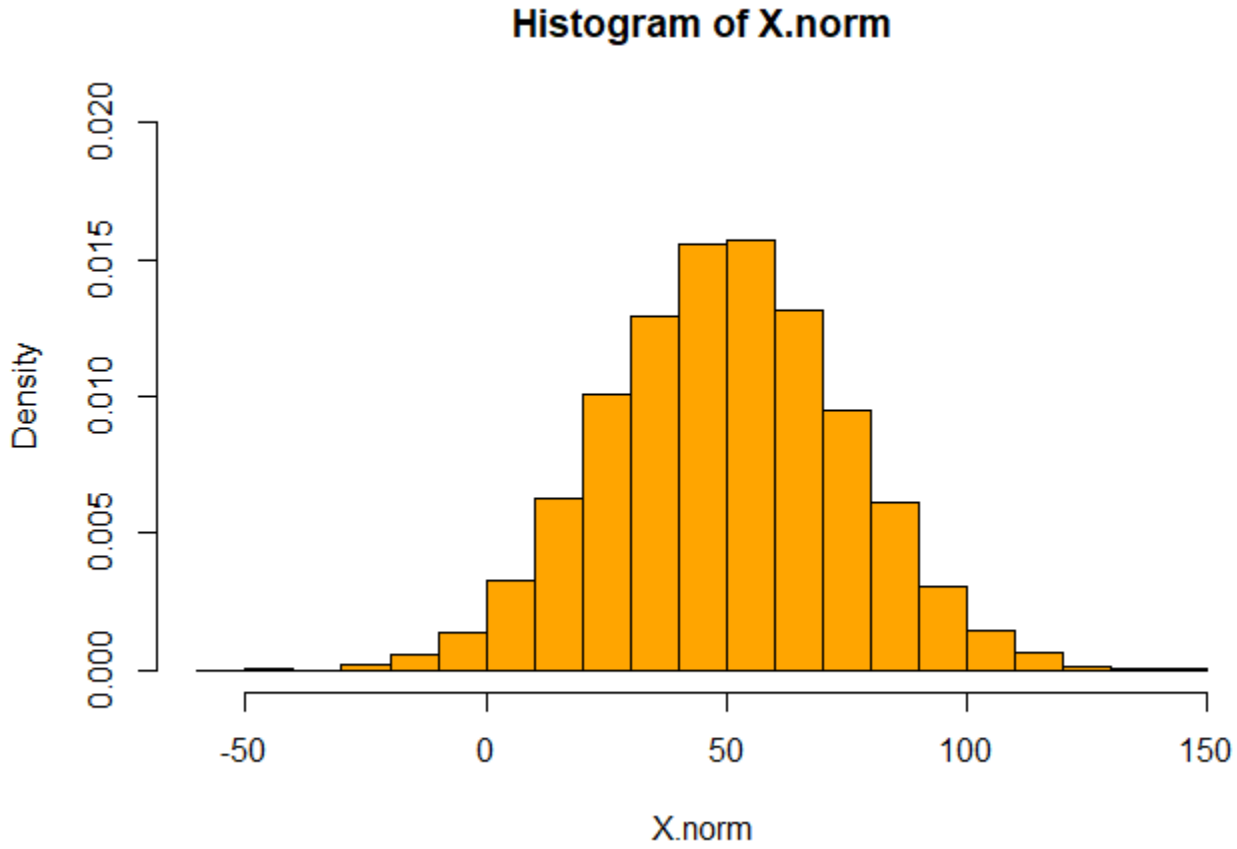
```
> x <- 1:9

> sample(x, size = 7)
[1] 9 2 6 7 4 5 1

> sample(x, size = 10)
Error in sample.int(length(x), size, replace, prob) :
  cannot take a sample larger than the population when 'replace = FALSE'

> sample(x, size = 10, replace = T)
[1] 1 8 8 7 9 5 6 4 7 8
```

중심극한정리: *central limit theorem*

- 표본의 크기가 충분히 클 때($n \geq 30$)
  - 표본분포는 모집단의 분포와 상관없이 정규분포를 따른다.
- 평균이 μ, 표준편차가 σ인 모집단으로부터 n개의 표본을 추출하면
  - 표본평균 $\bar{X}$의 확률분포는 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$인 정규분포에 근사한다.
  - 표본분포의 평균은 모집단의 평균 μ와 같다.
  - 표본분포의 표준편차(=표준오차)는 $\frac{\sigma}{\sqrt{n}}$와 같다.

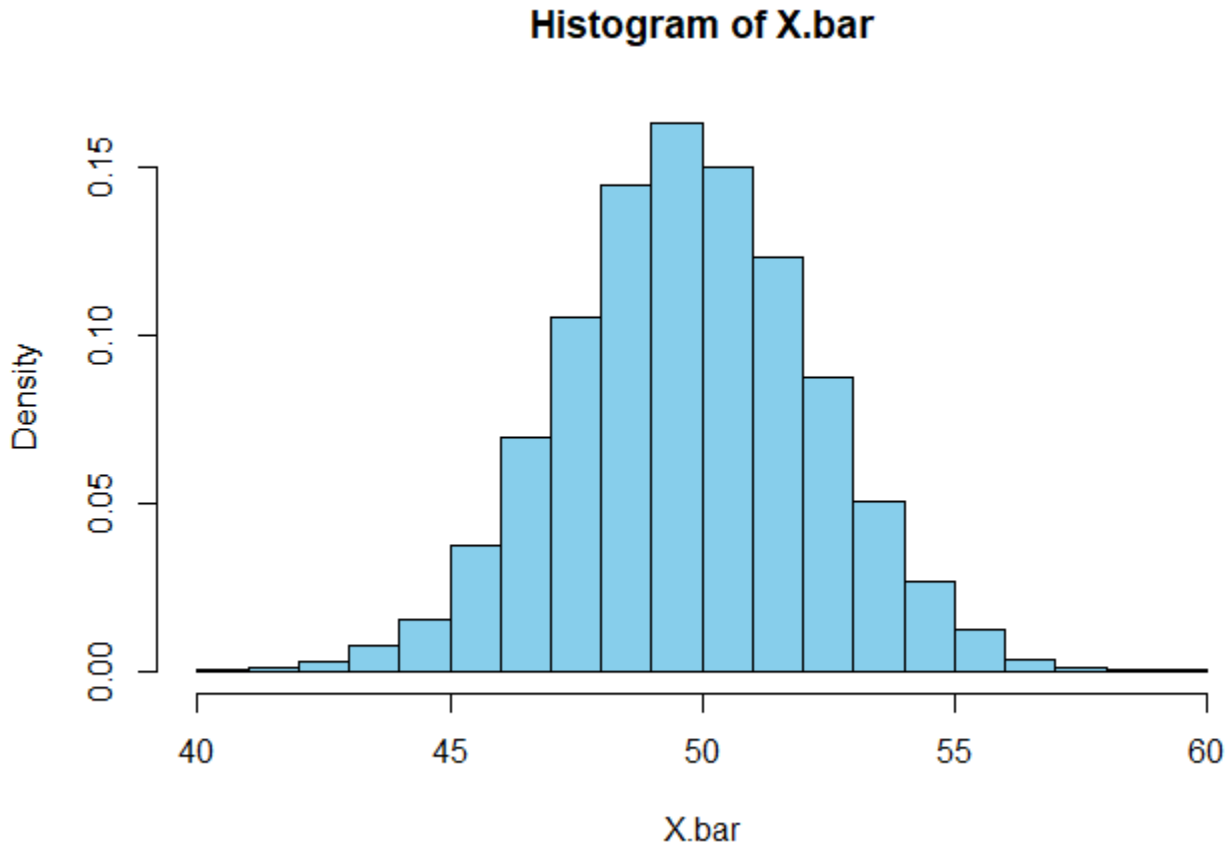
04. 확률변수와 확률분포 (2)

```
> X.norm <- rnorm(n = 10000, mean = 50, sd = 25)
> hist(X.norm, col = "orange", freq = F, ylim = c(0, 0.02))
> mean(X.norm)
[1] 49.73418
> sd(X.norm)
[1] 24.94434
```



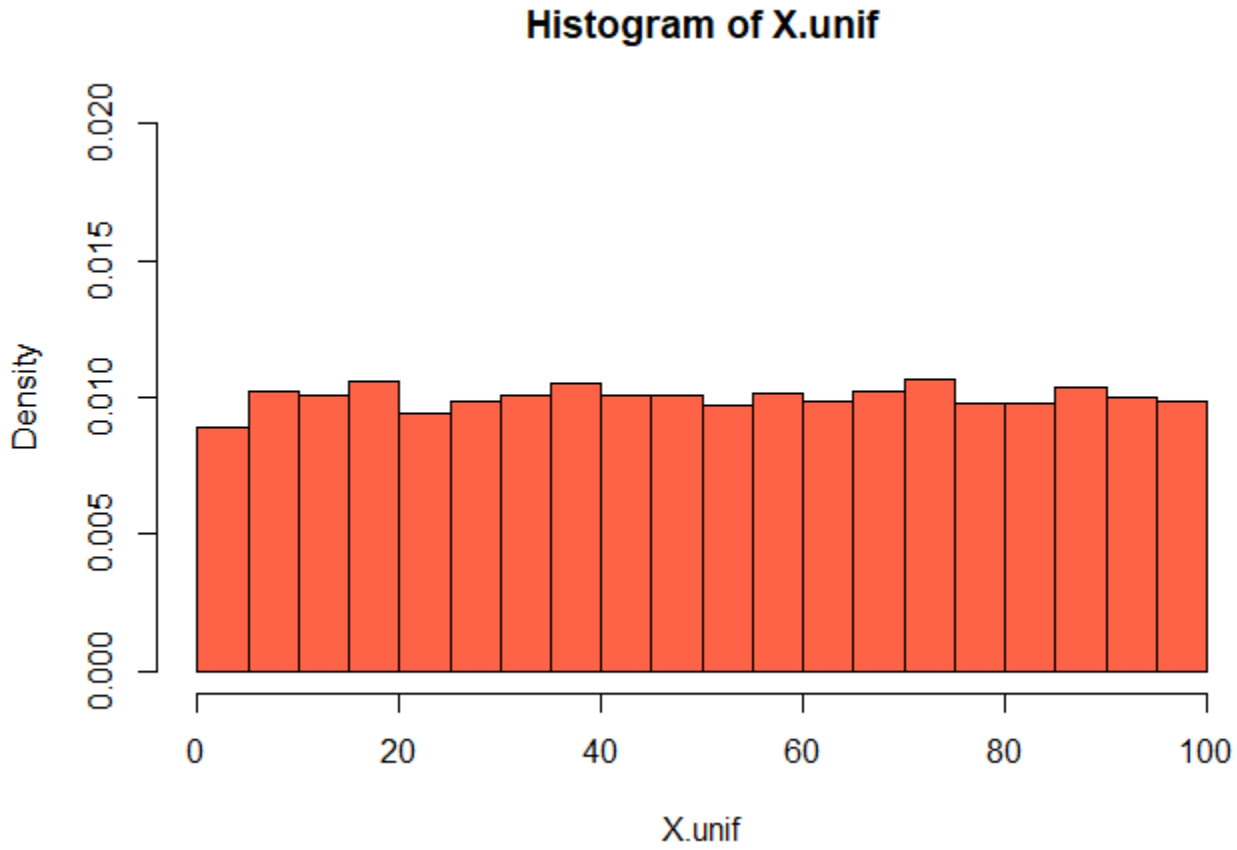
04. 확률변수와 확률분포 (2)

```
> X.bar <- c()
> for (i in 1:10000) {
+     X.bar <- c(X.bar, mean(sample(X.norm, size = 100)))
+ }
> hist(X.bar, col = "skyblue", freq = F)
```



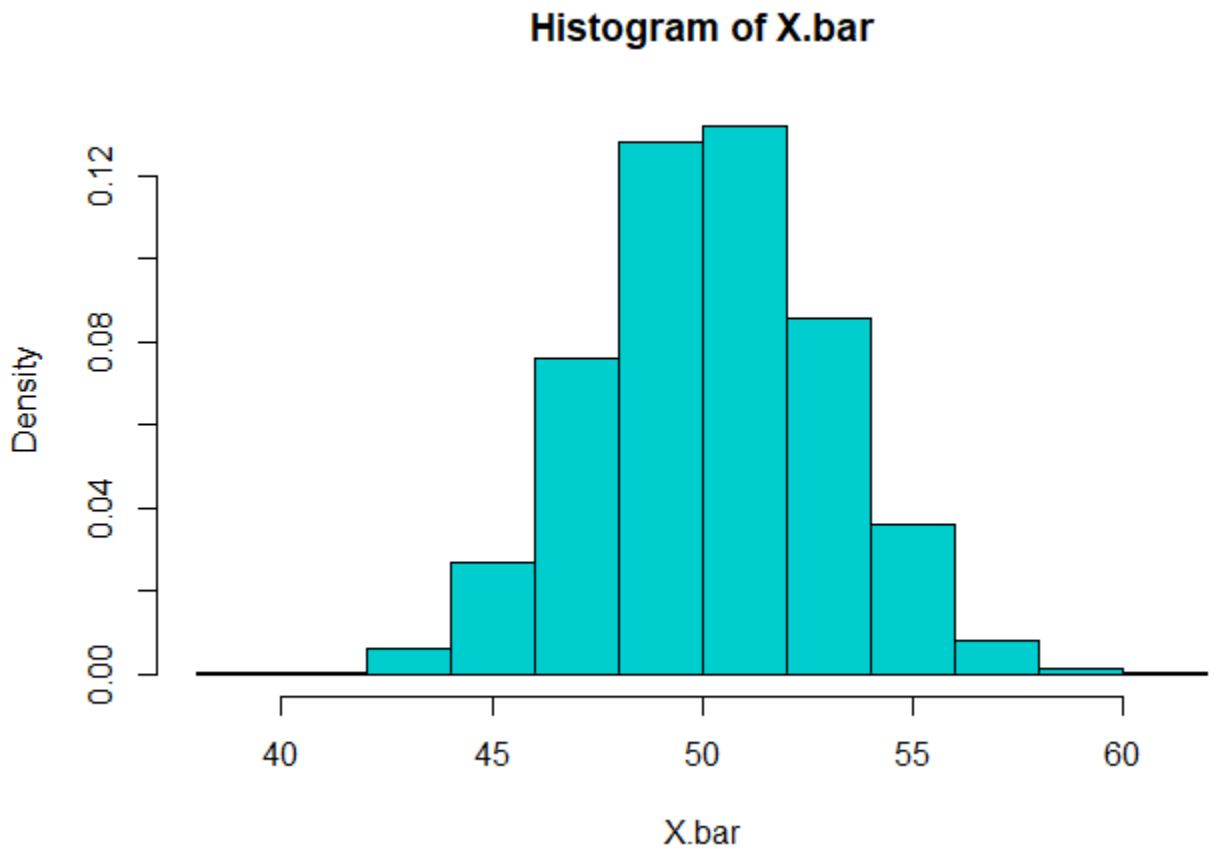
04. 확률변수와 확률분포 (2)

```
> X.unif <- runif(n = 10000, min = 0, max = 100)
> hist(X.unif, col = "tomato", freq = F, ylim = c(0, 0.02))
> mean(X.unif)
[1] 50.22894
> sd(X.unif)
[1] 28.6951
```



04. 확률변수와 확률분포 (2)

```
> X.bar <- c()
> for (i in 1:10000) {
+     X.bar <- c(X.bar, mean(sample(X.unif, size = 100)))
+ }
> hist(X.bar, col = "cyan3", freq = F)
```



Any Questions?