Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

05

통계적 추정과 가설검정

경북대학교 배준현 교수
(joonion@knu.ac.kr)

▪ 통계적 추정과 구간추정:

- 통계적 추정: 수집한 표본집단으로부터 모집단의 특성(모수)을 추정
  - 점추정: 모수에 대한 추정값을 하나의 값으로 추정
  - 구간추정: 모수의 값이 포함되리라고 믿을 수 있는 범위를 추정
- 구간추정의 신뢰수준과 신뢰구간
  - 신뢰수준: *confidence level*
    - 모수가 추정한 구간 안에 있을 것이라 믿을 수 있는 정도 (95%, 99%)
  - 신뢰구간: *confidence interval*
    - 신뢰도에 따라 모수가 포함될 것이라 믿을 수 있는 구간

▪ 정규분포와 신뢰구간:

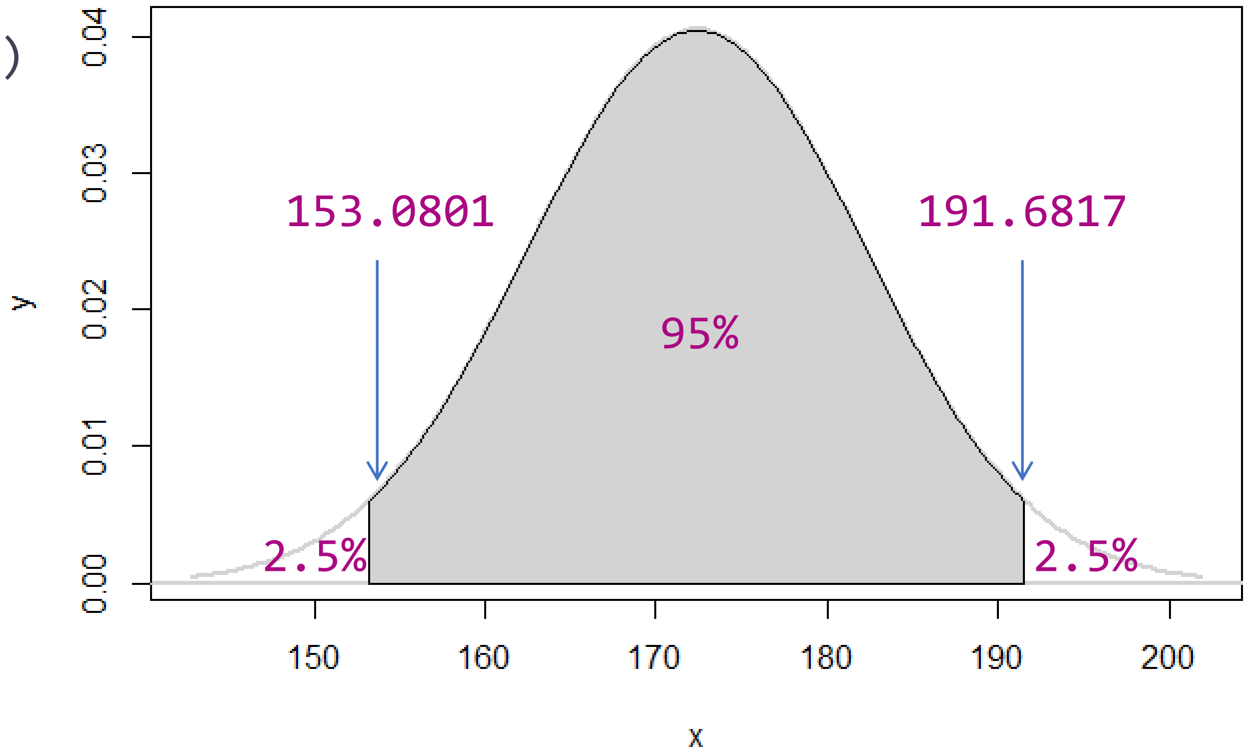
- 정규분포 $X \sim N(\mu, \sigma^2)$을 따르는 모집단에서
  - 크기가 n인 표본을 임의추출하여 얻은 표본평균이 $\bar{X}$일 때
- 모평균 μ에 대한 구간추정:
  - 신뢰수준 95%의 신뢰구간: $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$
  - 신뢰수준 99%의 신뢰구간: $\bar{X} - 2.58\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.58\frac{\sigma}{\sqrt{n}}$

```
> height <- survey$Height
> h.mean <- mean(height, na.rm = T)
> h.mean
[1] 172.3809
> h.sd <- sd(height, na.rm = T)
> h.sd
[1] 9.847528
> c(h.mean - 1.96*h.sd, h.mean + 1.96*h.sd)
[1] 153.0797 191.6820
```

05. 통계적 추정과 가설 검정

```
> qnorm(0.025, mean = h.mean, sd = h.sd)
[1] 153.0801
> qnorm(0.975, mean = h.mean, sd = h.sd)
[1] 191.6817
> pnorm(153.0801, mean = h.mean, sd = h.sd)
[1] 0.02500023
> pnorm(191.6817, mean = h.mean, sd = h.sd, lower.tail = F)
[1] 0.02499977
```

```
x <- seq(h.mean-3*h.sd, h.mean+3*h.sd, length.out = 200)
y <- dnorm(x, h.mean, h.sd)
plot(x, y, type = 'l', col = "lightgrey", lwd = 2)
abline(0, 0, col = "lightgrey", lwd = 2)
xlim <- x[lower <= x & x <= upper]
ylim <- y[lower <= x & x <= upper] - 0.0001
xlim <- c(xlim[1], xlim, tail(xlim, 1))
ylim <- c(0, ylim, 0)
polygon(xlim, ylim, col = "lightgrey")
```



- 가설검정: *hypothesis test*
  - 가설: *hypothesis*
    - 타당성의 유무를 명백히 밝혀야 하는 모수에 대한 주장
  - 가설검정: *hypothesis test*
    - 표본집단의 통계량을 이용하여 모수에 대한 주장의 진위를 검정하는 과정
  - 가설검정의 방법:
    - 연구가설에 대한 귀무가설과 대립가설을 설정하고,
    - 표본으로부터 얻은 검정통계량이 “귀무가설이 옳다”는 전제하에서는
      - 확률적으로 나타나기 어려운 극단적이고 예외적인 값이라는 것을 입증

- 귀무가설과 대립가설:
  - 연구주제: 경북대 대학원생의 평균 키가 한국 성인의 평균 키와 다를까?
  - 귀무가설 (H_0): *null hypothesis*
    - 모수에 대한 기존의 주장: 거짓으로 판단(기각)될 때까지 참으로 인정
    - 예) 경북대 대학원생의 평균 키는 한국 성인의 평균 키와 같다.
  - 대립가설 (H_1): *alternative hypothesis*
    - 모수에 대한 새로운 주장: 귀무가설이 거짓이라면 참(채택)이 되는 가설
    - 예) 경북대 대학원생의 평균 키는 한국 성인의 평균 키와 다르다.

▪ 가설검정의 오류:

- 제1종 오류: Type I error
  - 귀무가설이 참이지만, 검정 결과에 따라 귀무가설을 기각하는 오류(α)
- 제2종 오류: Type II error
  - 귀무가설이 거짓이지만, 검정 결과에 따라 귀무가설을 채택하는 오류 (β)

가설검정에 따른 판단

귀무가설의 실제 상황		H_0를 채택	H_0를 기각
	H_0가 참	-	1종 오류 (α = 유의수준)
	H_0가 거짓	2종 오류 (β)	검정력 ($1-\beta$)

- **통계적 유의성**: *statistical significance*
  - 유의수준(α): *significance level*
    - 1종 오류를 범할 통계적 확률
    - 일반적으로 표본으로부터 관측된 결과가 나타날 가능성이 5% 미만인 경우
      - 통계적으로 유의하다: 이러한 관측결과가 나타날 확률이 매우 낮다.
  - 유의확률(*p-value*): *significance probability*
    - 표본에서 관측한 통계량보다 더 극단적인 값이 발생할 확률
    - *p-value*가 α보다 크다: 귀무가설을 기각할 수 있는 증거가 부족함
    - *p-value*가 α보다 작다: 귀무가설을 기각할 수 있는 증거가 충분함

- **가설검정의 절차:**
  - 연구주제에 대한 귀무가설과 대립가설을 설정한다.
  - 유의수준 α를 결정한다. $(\alpha = 0.05)$
  - 적절한 검정통계량을 선택한다.
  - 표본으로부터 검정통계량의 유의확률 p-$value$를 구한다.
  - 귀무가설과 대립가설의 기각/채택 여부를 결정한다.
    - $p > \alpha$: 귀무가설을 기각할 수 없음
    - $p < \alpha$: 귀무가설을 기각하고 대립가설을 채택할 수 있음

이항분포와 가설검정

- 어떤 동전을 100번 던졌더니, 앞면이 60번 나왔다. $(p = 0.6)$
  - 이 동전은 조작이 없는 공평한 동전이라고 할 수 있을까?
- 귀무가설과 대립가설 설정
  - 귀무가설: 이 동전은 공평한 동전이 맞다. $(p = 0.5)$
  - 대립가설: 이 동전은 공평한 동전이 아니다. $(p \neq 0.5)$
- 유의수준 설정: $\alpha = 0.05$
  - 이항분포 $B(100, 0.5)$에서 성공확률이 $p = 0.6$인 결과가 나올 확률이
  - 신뢰수준 95% 범위 안에 있는가, 아니면 범위 밖에 있는가?
- 유의확률 확인: p-value
  - p-value < 0.05: 귀무가설을 기각하고 대립가설을 채택할 수 있음
  - p-value > 0.05: 귀무가설을 기각하고 대립가설을 채택할 수 없음

- 100번의 동전을 던져서 앞면이 60번 나왔다면, 공평한 동전이라고 할 수 있는가?

  H_0: 이 동전은 공평한 동전과 다르지 않다(성공확률이 $p = 0.5$이다).

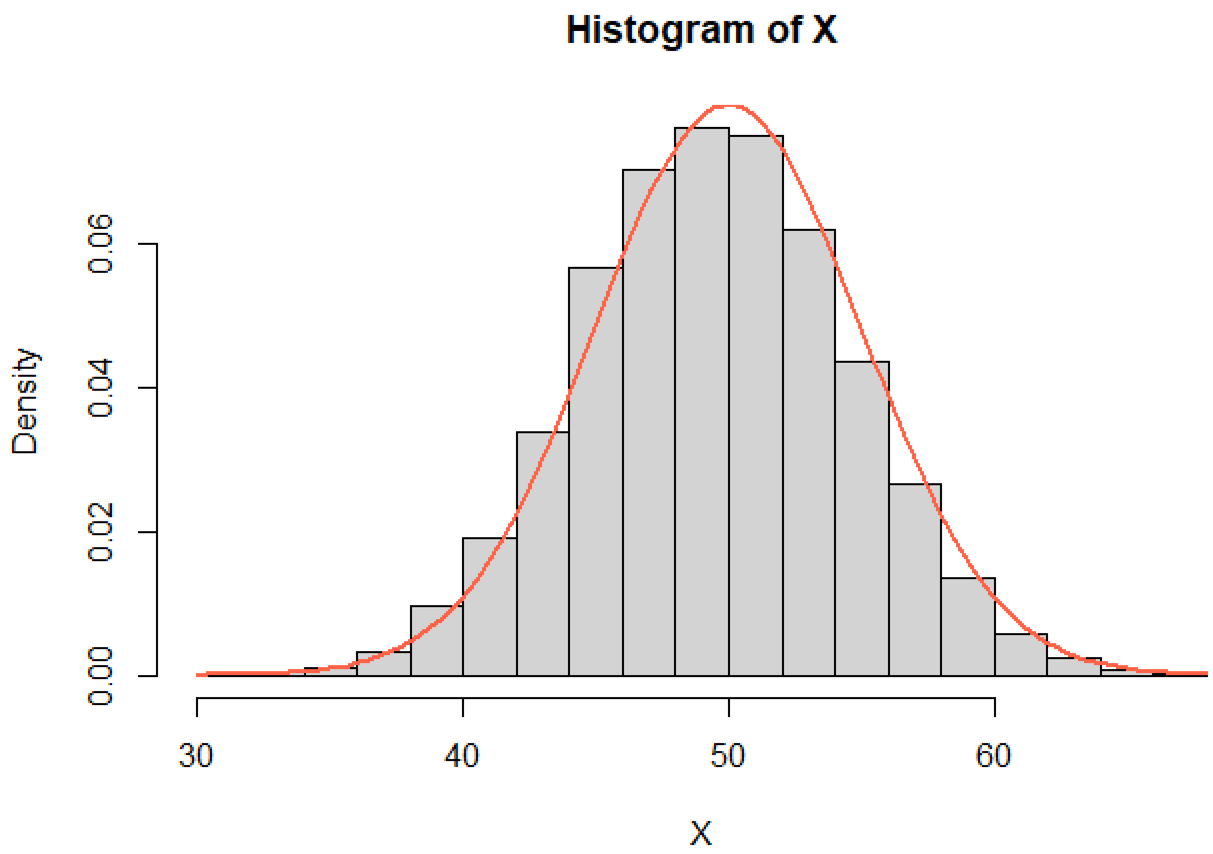
  H_1: 이 동전은 공평한 동전과 다르다(성공확률이 $p \neq 0.5$이다).

```
> binom.test(x = 60, n = 100, p = 0.5)

        Exact binomial test

data:  60 and 100
number of successes = 60, number of trials = 100, p-value = 0.05689
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4972092 0.6967052
sample estimates:
probability of success
                   0.6
```

p-value>0.05: 95%의 신뢰구간 안쪽에 $p = 0.5$가 포함됨.

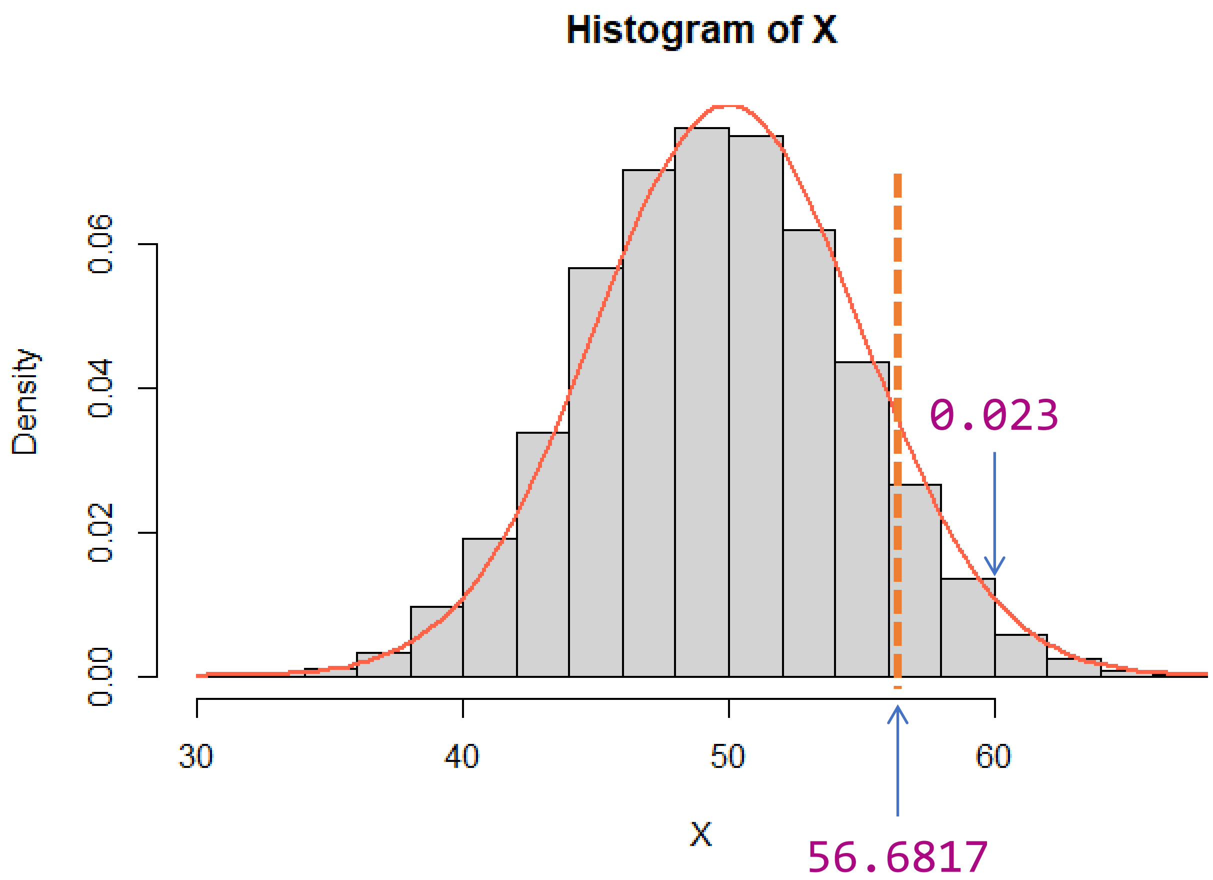
따라서, 유의수준 $\alpha = 0.05$에서 귀무가설을 기각할 수 없다.

```
> set.seed(2022)
> X <- rbinom(n = 10000, size = 100, prob = 0.5)
> hist(X, col = "lightgray", breaks = 15, freq = F)
> x <- seq(0, 100, 1)
> curve(dnorm(x, mean(X), sd(X)), add = T, col = "tomato", lwd = 2)
```



05. 통계적 추정과 가설 검정

```
> qnorm(p=0.095, mean(X), sd(X), lower.tail = F)
[1] 56.56878

> pnorm(q=60, mean(X), sd(X), lower.tail = F)
[1] 0.0231243
```



- 100번의 동전을 던져서 앞면이 65번 나왔다면, 공평한 동전이라고 할 수 있는가?
  H_0: 이 동전은 공평한 동전과 다르지 않다(성공확률이 $p = 0.5$이다).
  H_1: 이 동전은 공평한 동전과 다르다(성공확률이 $p \neq 0.5$이다).

```
> binom.test(x = 65, n = 100, p = 0.5)

	Exact binomial test

data:  65 and 100
number of successes = 65, number of trials = 100, p-value = 0.003518
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5481506 0.7427062
sample estimates:
probability of success 
                  0.65 
```

p-value<0.05: 95%의 신뢰구간 바깥쪽에 $p = 0.5$가 위치함.
따라서, 유의수준 $\alpha = 0.05$에서 귀무가설을 기각할 수 있다.

- 100번의 동전을 던져서 앞면이 35번 나왔다면, 공평한 동전이라고 할 수 있는가?

  H_0: 이 동전은 공평한 동전과 다르지 않다(성공확률이 $p = 0.5$이다).

  H_1: 이 동전은 공평한 동전과 다르다(성공확률이 $p \neq 0.5$이다).

  단, 유의수준을 $\alpha = 0.01$로 결정 (신뢰수준 99%)

```
> binom.test(35, 100, 0.5, conf.level = 0.99)

        Exact binomial test

data:  35 and 100
number of successes = 35, number of trials = 100, p-value = 0.003518
alternative hypothesis: true probability of success is not equal to 0.5
99 percent confidence interval:
 0.2319154 0.4827790
sample estimates:
probability of success
                  0.35
```

- **양측검정과 단측검정:**
  - 양측검정: *two-sided* hypothesis test
    - 귀무가설 $H_0: \theta = \theta_0$에 대해 대립가설을 $H_0: \theta \neq \theta_0$로 설정
  - 단측검정: *one-sided* hypothesis test
    - 하단측검정: one-sided *lower* hypothesis test
      - 귀무가설 $H_0: \theta \geq \theta_0$에 대해 대립가설을 $H_1: \theta < \theta_0$로 설정
    - 상단측검정: one-sided *upper* hypothesis test
      - 귀무가설 $H_0: \theta \leq \theta_0$에 대해 대립가설을 $H_1: \theta > \theta_0$로 설정

05. 통계적 추정과 가설 검정

- 100번의 동전을 던져서 앞면이 60번 나왔다면, 이 동전은 앞면이 더 많이 나오는 동전일까?

  H_0: 이 동전은 앞면이 더 많이 나오지 않는다(성공확률이 $p \leq 0.5$이다).

  H_1: 이 동전은 앞면이 더 많이 나온다(성공확률이 $p > 0.5$이다).

```
> binom.test(60, 100, 0.5, alternative = "greater")

	Exact binomial test

data:  60 and 100
number of successes = 60, number of trials = 100, p-value = 0.02844
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.5129758 1.0000000
sample estimates:
probability of success 
                   0.6
```

p-value<0.05: 95%의 신뢰구간 바깥쪽에 $p = 0.5$가 위치함.

따라서, 유의수준 $\alpha = 0.05$에서 귀무가설을 기각할 수 있다.

- 100번의 동전을 던져서 앞면이 45번 나왔다면, 이 동전은 앞면이 더 적게 나오는 동전일까?
  H_0: 이 동전은 앞면이 더 적게 나오지 않는다(성공확률이 $p \geq 0.5$이다).
  H_1: 이 동전은 앞면이 더 적게 나온다(성공확률이 $p < 0.5$이다).

```
> binom.test(45, 100, 0.5, alternative = "less")

	Exact binomial test

data:  45 and 100
number of successes = 45, number of trials = 100, p-value = 0.1841
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.5371104
sample estimates:
probability of success 
                  0.45 
```

p-value>0.05: 95%의 신뢰구간 안쪽에 $p = 0.5$가 위치함.
따라서, 유의수준 $\alpha = 0.05$에서 귀무가설을 기각할 수 없다.

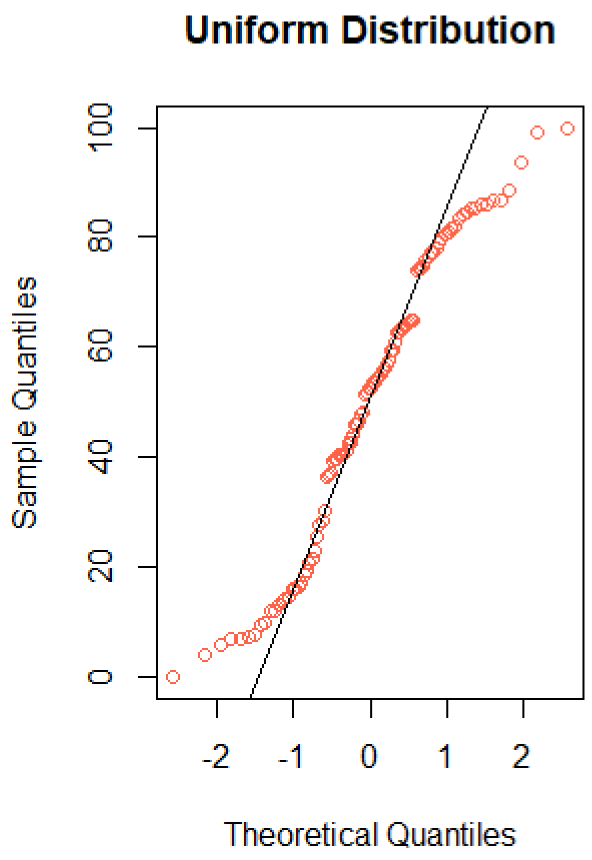
- 정규성 검정: *normality test*
  - 정규성 가정: 통계분석의 여러 검정방법들이 데이터가 정규분포임을 가정
    - 정규성을 가정하는 통계분석 방법을 사용할 때는 정규성을 검정해야 함
  - 정규성 검정: *normality test*
    - 귀무가설: 데이터의 분포가 정규분포를 따른다.
    - 대립가설: 데이터의 분포가 정규분포를 따르지 않는다.
  - 샤피로-윌크 검정: *Shapiro-Wilk* normality test
    - 귀무가설: 표본 데이터가 정규성을 만족한다.
      - 유의수준 0.05를 적용하면,
      - p-value가 0.05보다 클 때 정규성을 만족한다고 주장할 수 있음
    - `shapiro.test(x)`
      - `x`: a numeric vector of data values.

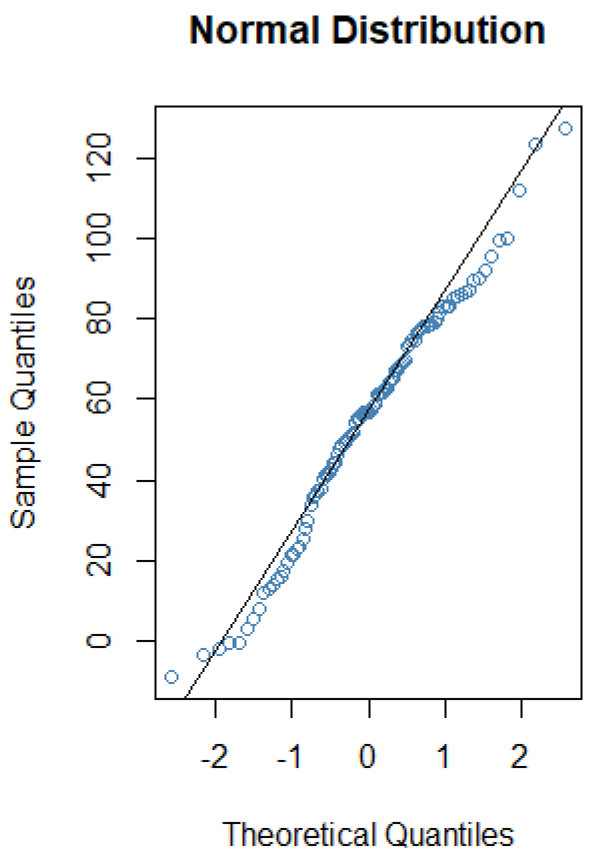
- survey 데이터셋의 Height와 Age는 각각 정규성을 만족하는가?

```
> shapiro.test(survey$Height)

        Shapiro-Wilk normality test

data:  survey$Height
W = 0.98841, p-value = 0.08844

> shapiro.test(survey$Age)

        Shapiro-Wilk normality test

data:  survey$Age
W = 0.45642, p-value < 2.2e-16
```

```
> set.seed(2022)
> x.unif <- runif(100, min = 0, max = 100)
> x.norm <- rnorm(100, mean(x.unif), sd(x.unif))

> shapiro.test(x.unif)

	Shapiro-Wilk normality test

data:  x.unif
W = 0.95553, p-value = 0.001954

> shapiro.test(x.norm)

	Shapiro-Wilk normality test

data:  x.norm
W = 0.98319, p-value = 0.2336
```

```r
par(mfrow = c(1, 2))
qqnorm(x.unif, col = "tomato", main = "Uniform Distribution")
qqline(x.unif)
qqnorm(x.norm, col = "steelblue", main = "Normal Distribution")
qqline(x.norm)
par(mfrow = c(1, 1))
```


Uniform Distribution
Sample Quantiles
Theoretical Quantiles
0
20
40
60
80
100
-2
-1
0
1
2


Normal Distribution
Sample Quantiles
Theoretical Quantiles
0
20
40
60
80
100
120
-2
-1
0
1
2

Any Questions?