

강의교안 이용 안내

- 본 강의교안의 저작권은 이윤환과 한빛아카데미(주)에 있습니다.
- 이 자료를 무단으로 전제하거나 배포할 경우 저작권법 136조에 의거하여 벌금에 처할 수 있고 이를 병과(併科)할 수도 있습니다.





제대로 알고 쓰는
R 통계분석

CHAPTER 05

추정

Contents

1.1

점추정

- 통계학이란
- 모집단과 표본, 그리고 기본원리
- 통계에서의 자료

1.2

구간추정

- R이란
- R 기초

6장을 위한 준비



01. 점추정

: 모수의 값을 점으로 추측하기

1. 추측통계학의 개념과 점추정 구간추정의 특성을 이해한다.
2. 좋은 추정량이 갖춰야 하는 성질에 대해 학습한다.
3. 표준오차의 의미에 대해 학습한다.

추측통계학

- **추측통계학**

- 표본으로부터 특성을 관찰하여 모집단의 특성을 유추하는 통계학의 한 분야
- 추측통계학의 두 가지 연구 방법
 - 추정 : 모집단으로부터 추출된 표본으로부터 특성(통계량)을 파악하여 이를 바탕으로 모수를 유추하는 방법
 - 가설검정 : 모수에 대한 가설을 수립하고 이로부터 어떤 가설을 선택할 것인지를 통계적으로 결정하는 방법
- 추정의 종류
 - 점추정 : 표본의 특성을 나타내는 계산식(통계량) 중 모수를 유추하는 데 있어 최적의 계산식을 통해 구한 하나의 추정값을 구하는 방법
 - 점추정은 표본으로부터 계산되는 값이기에 추출되는 표본에 따라 값이 달라진다.
 - 구간추정 : 하나의 점(값)이 아닌 모수의 참값이 포함될 것으로 기대하는 구간을 추정하는 방법입니다

추정량

• 추정량

- 알고자 하는 모수(θ)를 추측하기 위해 표본으로부터 관찰된 값으로 계산되는 표본의 통계량으로 $\hat{\theta}$ 으로 표기합니다.
- 추정치 : 하고, 표본으로부터 관측된 자료를 통해 계산된 추정량의 결과(값)
- 모수와 추정량

모수(θ)	구분	추정량($\hat{\theta}$)
μ	평균	\bar{X}
σ^2	분산	s^2
P	비율	\hat{p}

좋은 추정량

- 불편성과 불편추정량

- 불편성

- 추정량이 갖춰야 할 가장 기본적인 성질로 한쪽으로 치우쳐지지 않음을 의미합니다.
 - ‘치우쳐지지 않음’은 추정량의 기대값이 모수와 같음을 나타내며 이런 성질을 만족하는 추정량을 불편추정량이라고 합니다.

불편추정량

모수 θ 에 대한 추정량 $\hat{\theta}$ 이 다음을 만족할 때 $\hat{\theta}$ 은 θ 에 대한 불편추정량이라 합니다.

$$E(\hat{\theta}) = \theta$$

좋은 추정량

▣ 불편추정량 판정 : 표본평균

- 평균이 μ 이고 분산이 σ^2 인 모집단으로 부터 추출한 n 개의 확률표본을 X_1, X_2, \dots, X_n

이라 할 때 표본평균 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 가 불편추정량인지 확인해 봅시다. (부록 D의 기대값의 성질 참고)

- $E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i)$

- X_i 들의 기대값은 모집단의 평균인 μ 이며 μ 는 알지 못할 뿐 존재하는 상수입니다.

- $\frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \frac{1}{n} \cdot n \cdot \mu = \mu$

- 따라서 $E(\bar{X}) = \mu$ 가 되어 표본평균은 불편추정량입니다.

- ▣ 추정량의 기대값이 모수와 같음

좋은 추정량

• 유효성

- 모수에 대한 불편추정량은 한가지 이상 존재할 수 있습니다.
- 여러 개의 불편추정량이 있을 때 좋은 추정량을 결정하는 성질입니다.

더 유효한 추정량

모수 θ 에 대한 두 불편추정량 $\widehat{\theta}_1, \widehat{\theta}_2$ 에 대해 각각의 분산을 $Var(\widehat{\theta}_1), Var(\widehat{\theta}_2)$ 라 할 때, 다음을 만족하면 $\widehat{\theta}_1$ 이 $\widehat{\theta}_2$ 보다 ‘더 유효한 추정량’이라고 합니다.

$$Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2)$$

- 예제) 평균이 μ 이고 분산이 σ^2 인 모집단으로 부터 독립으로 추출한 3개의 확률표본을 X_1, X_2, X_3 라 할 때, 다음과 같은 모평균에 대한 두 개의 추정량 \bar{Y}_1, \bar{Y}_2 이 있을 경우 더 유효한 추정량은 어떤 것일지 판단해 봅시다.

$$\bar{Y}_1 = \frac{X_1 + X_2 + X_3}{3}, \quad \bar{Y}_2 = \frac{X_1 + 2X_2 + 3X_3}{6}$$

좋은 추정량

- 먼저 두 추정량이 불편추정량인지 확인해 봅시다.

- \bar{Y}_1 는 표본 3개의 표본평균이므로 불편추정량입니다.

- $E(\bar{Y}_2) = E\left(\frac{X_1+2X_2+3X_3}{6}\right) = \frac{1}{6}E(X_1 + 2X_2 + 3X_3) = \frac{1}{6}(\mu + 2\mu + 3\mu) = \frac{1}{6}6\mu = \mu$

- \bar{Y}_2 역시 불편추정량입니다.

- 두 추정량의 분산을 구해봅시다.

- $Var(\bar{Y}_1) = Var\left(\frac{X_1+X_2+X_3}{3}\right) = \frac{1}{9}Var(X_1 + X_2 + X_3) = \frac{1}{9}(\sigma^2 + \sigma^2 + \sigma^2) = \frac{1}{3}\sigma^2$

- $Var(\bar{Y}_2) = Var\left(\frac{X_1+2X_2+3X_3}{6}\right) = \frac{1}{36}Var(X_1 + 2X_2 + 3X_3) = \frac{1}{36}(\sigma^2 + 4\sigma^2 + 9\sigma^2) =$

$$\frac{14}{36}\sigma^2 = \frac{7}{18}\sigma^2$$

- 이로부터 $Var(\bar{Y}_1) < Var(\bar{Y}_2)$ 이며 \bar{Y}_1 이 \bar{Y}_2 보다 더 유효한 추정량입니다.

좋은 추정량

예제 5-2 유효성 모의실험

준비파일 | 02.efficiency_simulation.R

- 앞서 사용한 두 추정량 \bar{Y}_1, \bar{Y}_2 의 분포를 확인해 봅시다.
 - 표준정규분포를 이루는 모집단에서 3개의 확률표본을 추출합니다.
 - \bar{Y}_1 는 mean() 함수를 이용하여 구하고 \bar{Y}_2 는 새로운 함수를 만들어 구합니다.

```
3. mean.seq <- function (x) {  
4.   n <- length(x)  
5.   sum <- 0  
6.   n2 <- 0  
7.   for( i in 1:n) {  
8.     newx <- i * x[i]  
9.     sum <- sum + newx  
10.    n2 <- n2 + i  
11.  }  
12.  return( sum / n2 )  
13.}
```

좋은 추정량

- **Step #1) \bar{Y}_2 를 계산하기 위한 함수 `mean.seq()`를 생성합니다.**
 - 3줄 : `mean.seq`는 한 개의 자료를 `x`라는 이름으로 전달받아 사용합니다.
 - 전달받은 `x`는 \bar{Y}_2 를 계산할 표본들의 벡터입니다.
 - 4줄 : 전달받은 벡터 `x`의 원소 수, 즉 표본의 개수를 변수 `n`에 저장합니다.
 - 5줄, 6줄 : \bar{Y}_2 계산을 위한 분모와 분자의 합을 구하기 위해 합의 항등원으로 초기화 합니다.
 - 분자에 해당하는 표본별 합을 구하기 위한 변수 `sum`을 0으로 초기화
 - 분모는 벡터 `x`의 각 값에 곱해지는 값들의 합으로 이를 위해 변수 `n2`를 0으로 초기화
 - 7, 11줄 : 전달받은 벡터 `x`의 각 값을 원소별로 접근하기 위해 반복문을 사용합니다.
 - 8줄 : 벡터 `x`의 `i`번째 원소(`x[i]`)와 `i`를 곱해 변수 `newx`에 저장합니다.
 - 9줄 : 위에서 구한 `newx`와 기존에 있던 `sum`과 합한 값으로 변수 `sum`의 값을 변경합니다.
 - 10줄 : `sum`의 경우와 마찬가지로 `i`와 기존에 있던 `n2`의 값을 합한 값으로 변수 `n2`의 값을 변경합니다.
 - 12줄 : 위에서 구한 `sum`을 `n2`로 나눈 값, 즉 $\bar{Y}_2 = \frac{X_1+2X_2+3X_3}{6}$ 를 반환합니다.

좋은 추정량

```
15. y1 <- rep(NA, 1000)
16. y2 <- rep(NA, 1000)
17. for(i in 1:1000) {
18.   smp <- rnorm(3)
19.   y1[i] <- mean(smp)
20.   y2[i] <- mean.seq(smp)
21. }
```

- **Step #2)** 표준정규분포로부터 3개씩의 표본을 뽑아 \bar{Y}_1, \bar{Y}_2 를 구하는 과정을 1,000번 반복합니다.
 - 15줄, 16줄 : 1000번의 표본추출로 구해지는 \bar{Y}_1, \bar{Y}_2 를 저장하기 위한 변수 y1, y2를 준비합니다.
 - rep() 함수로 결측값(NA) 1,000개를 원소로 갖는 벡터를 생성하고 y1과 y2에 저장
 - 17, 21줄 : 표본추출을 1,000번 실시하기 위해 반복문을 사용합니다.
 - 18줄 : 표준정규분포로부터 3개의 표본을 추출하고, 이를 변수 smp에 저장합니다.
 - 19줄 : 위에서 추출한 3개의 표본의 평균을 구해 y1의 i번째에 저장합니다.
 - 20줄 : 위에서 추출한 3개의 표본의 \bar{Y}_2 를 구해 y2의 i번째에 저장합니다.

좋은 추정량

```
23. n1 <- length(y1[(y1 > -0.1) & (y1 < 0.1)])
24. n2 <- length(y2[(y2 > -0.1) & (y2 < 0.1)])
25. data.frame(mean=mean(y1), var=var(y1), n=n1)
26. data.frame(mean=mean(y2), var=var(y2), n=n2)
```

- **Step #3) 결과를 확인합니다.**

- 23줄 : 3개로 구성된 표본의 평균 1,000개가 저장된 y1에서 그 값이 (모평균 주변인) -0.1보다 크고 0.1보다 작게 나온 횟수를 구해 변수 n1으로 저장합니다.
- 24줄 : \bar{Y}_2 의 값 1,000개가 저장된 y2에서 그 값이 (모평균 주변인) -0.1보다 크고 0.1보다 작게 나온 횟수를 구해 변수 n2로 저장합니다.
- 25, 26줄 : 모의실험한 두 추정량 평균과 각각의 평균(mean), 분산(var), 그리고 위에서 구한 -0.1부터 0.1사이에 있는 값의 개수(n)를 출력합니다.
 - 추정량으로 표본평균을 사용한 경우(n1)의 개수가 \bar{Y}_2 를 사용한 경우 (n2) 보다 많습니다. 즉, 모평균 주변에 좀더 많이 몰려 있습니다.

좋은 추정량

- 일치성

- 표본의 크기와 관련이 있는 성질입니다.
- 일치추정량

일치추정량

모수 θ 에 대한 추정량 $\hat{\theta}$ 라 할 때 임의의 양수 ε 에 대해 다음을 만족하면 $\hat{\theta}$ 은 θ 에 대한 일치추정량이라고 합니다.

$$\lim_{n \rightarrow \infty} P(|\hat{\theta} - \theta| > \varepsilon) = 0$$

- 표본의 크기가 커질수록 추정량의 추정치가 모수와 차이가 작아져 확률적으로 같아짐을 의미합니다.

표준오차

• 표준오차

▣ 4장을 다시 살펴 봅시다.

- 표본의 크기가 크다면, 중심극한정리에 의해 평균이 μ 이고 분산이 σ^2 인 임의의 분포로부터 추출된 확률표본 X_1, X_2, \dots, X_n 의 표본평균의 분포는 평균이 μ 이고 표준편차가 $\frac{\sigma}{\sqrt{n}}$ 인 정규분포($N(\mu, (\frac{\sigma}{\sqrt{n}})^2)$)를 따름을 학습했습니다.
- 표본평균들의 분포에서 기대값은 모집단의 평균과 동일함을 학습했습니다.

▣ 2장에서 살펴본 표준편차

- '자료들이 자신들의 평균을 중심으로 얼마나 퍼져있을지'를 나타내는 척도로 평균과 단위가 동일합니다.
- 표준편차가 작다면 표본평균들이 자신들의 평균인 주변에 많이 모여 있을 것으로 기대되고, 그 값이 크다면 평균 주변에 몰려있기보다는 전체적으로 많이 퍼져있을 것으로 기대됨을 학습했습니다.

표준오차

- ▣ 앞서 학습한 내용을 바탕으로 다음을 생각해 봅시다.
 - 표본평균들의 분포에서 표본평균을 임의로 추출하는 경우입니다.
 - 표준편차가 작은 경우에는 모평균 μ 와 가까운 값이 나타날 가능성이 클 것으로 기대합니다.
 - 모평균과 임의로 추출한 표본평균의 차이가 크지 않을 것으로 여겨집니다.
 - 표준편차가 큰 경우에는 모평균 μ 와 가까운 값이 나타날 가능성이 작을 것으로 기대합니다.
 - 모평균과 가까울 수도 있고, 멀 수도 있을 가능성이 크므로 모평균 주변에 있는지 판단하기 어려워집니다.
- ▣ 표준오차
 - 추정에서는 추정량의 표준편차에 대해 그 값이 작으면 모평균 추정에 대한 신뢰도가 높아지고 크면 신뢰도가 낮아지게 되므로 작을수록 좋은 개념인 오차를 사용합니다.

표준오차

▣ 표준오차 : $SE(\hat{\theta})$

- 모평균 추정에 사용한 추정량으로써의 표본평균의 표준오차는 다음과 같습니다.





$$SE(\hat{\theta}) = \frac{\sigma}{\sqrt{n}}$$

- 분모는 표본의 크기의 제곱근으로 표본에 의해 결정됩니다.
 - 제곱근이기에 표준오차를 반으로 줄으려면 표본은 4배가 더 필요합니다.
- 분자는 모집단의 표준편차로 우리가 알지 못하는 경우가 더 많습니다.
 - 즉, 모집단의 표준편차를 모르므로 계산할 수 없는 경우가 많습니다.
 - 표준오차 역시 하나의 모수로 우리가 추정해야 할 대상이 됩니다.
 - 표준오차의 추정은 모집단 표준편차 대신 표본으로부터 관찰하는 표본표준편차를 추정량으로 사용합니다.

$$\widehat{SE(\hat{\theta})} = \frac{s}{\sqrt{n}}$$

표준오차

- 좋은 추정량은 다음중 어떤 추정량일 까요?
 - 동심원은 과녁을 나타내어 중심이 모수입니다.

	큰 표준오차	작은 표준오차
편이된 추정량(불편 추정량이 아닌 경우)		
불편 추정량		

예제 : 모비율에 대한 점추정

- 모집단에서 원하는 결과가 나타날(성공할) 비율 P 에 대한 점추정을 실시해봅시다.
- 모비율 P 에 대한 추정량 : 표본비율 \hat{p}
 - 모집단으로부터 n 개의 확률 표본을 추출했을 때 원하는 결과의 개수(성공의 개수) X 의 비율입니다.

$$\hat{p} = \frac{X}{n}$$

- 여기서 성공의 개수 X 는 시행 횟수가 n 이고, 성공 확률이 모집단에서 원하는 결과가 나타날 비율 P 인 이항분포를 따르는 확률변수입니다($X \sim B(n, P)$)
- 표본비율 \hat{p} 의 기대값은 불편추정량입니다.(이항분포의 기대값 이용)
 - $E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n} \cdot nP = P$

예제 : 모비율에 대한 점추정

- 표본비율 \hat{p} 의 표준편차 : 표준오차

$$SE(\hat{p}) = \sqrt{Var(\hat{p})} = \sqrt{Var\left(\frac{X}{n}\right)} = \sqrt{\frac{1}{n^2} Var(X)} = \sqrt{\frac{1}{n^2} nP(1-P)} = \sqrt{\frac{nP(1-P)}{n}}$$

- 만약 모비율 P 를 알지 못하는 경우에는 $SE(\hat{p})$ 에서 사용한 모비율 P 의 추정량인 표

본비율 \hat{p} 을 사용하여 추정된 표준오차 $\widehat{SE}(\hat{p}) = \sqrt{\frac{n\hat{p}(1-\hat{p})}{n}}$ 로 구합니다.

예제 : 모비율에 대한 점추정을 위한 표본비율의 분포

예제 5-3 모비율에 대한 점추정

준비파일 | 03.prop.est.R

- 실습내용

- 주사위를 세 번 굴려 나오는 짝수의 비율을 표본비율로하여 표본비율들의 분포를 구해 봅시다.

```
1. library(prob)
2. n <- 3
3. smps.all <- rolldie(n)
4. str( smps.all )
5. head( smps.all, n=3 )
```

- **Step #1)** 주사위를 세 번 굴리면 나오는 눈의 수를 관찰하는 모든 경우의 수를 생성합니다.
 - 1줄 : rolldie() 함수를 사용하기 위해 3장에서 학습한 prob 패키지를 사용합니다.

예제 : 모비율에 대한 점추정을 위한 표본비율의 분포

- ▣ 2줄 : 주사위를 굴리는 횟수 3회를 변수 n 에 저장합니다.
- ▣ 3줄 : `rolldie()` 함수에 표본추출횟수가 저장된 변수 n 을 전달하여, 주사위를 세 번 굴릴 경우 관찰할 수 있는 모든 경우의 수를 `smps.all`에 데이터 프레임으로 저장합니다.
- ▣ 4줄 : 위에 저장한 데이터프레임은 X_1, X_2, X_3 세 개의 변수(열)로 구성된 데이터 프레임으로 2^{16} 개의 관찰치와 3개의 변수가 있습니다. 변수들은 각각 첫 번째, 두 번째, 세 번째 굴려서 나온 주사위의 눈의 값이 저장됩니다.
- ▣ 5줄 : 위의 저장된 `smps.all` 데이터 프레임의 앞 3개($n=3$)의 관찰치를 살펴
다

예제 : 모비율에 대한 점추정을 위한 표본비율의 분포

```
> str( smps.all )
'data.frame':  216 obs. of  3 variables:
 $ X1: int   1 2 3 4 5 6 1 2 3 4 ...
 $ X2: int   1 1 1 1 1 1 2 2 2 2 ...
 $ X3: int   1 1 1 1 1 1 1 1 1 1 ...
> head( smps.all, n=3 )
  X1 X2 X3
1  1  1  1
2  2  1  1
3  3  1  1
```

```
7. is.even <- function(x) return(!x%%2)
8. var.p <- function(x) {
9.   return( sum((x-mean(x))^2 / length(x))  )
10.}
11.p.even <- function(x, s.size=3) {
12.  return( sum(is.even(x)) / s.size )
13.}
```


예제 : 모비율에 대한 점추정을 위한 표본비율의 분포

- **Step #2)** 필요로 하는 함수들을 만듭니다.
 - 7줄 : 함수 생성 시 함수의 몸체가 한 줄 밖에 없을 경우 중괄호({ }) 없이 생성할 수 있습니다.
 - 생성하는 함수의 이름은 `is.even`이고, 이 함수는 전달인자 한 개를 받아 이를 변수 `x`에 저장하고 변수 `x`로 함수 내부에서 사용합니다.
 - 짝수는 `TRUE`, 홀수는 `FALSE`로 변환한 결과를 반환하는 함수입니다.
 - 8~10줄 : `var.p`는 전달인자 하나를 변수 `x`로 받아 모분산을 계산하고 그 값을 전달합니다.
 - 11~13줄 : `p.even` 함수는 두 개의 전달인자를 받도록 되어 있습니다.
 - 첫 번째 전달인자는 함수에서 사용할 데이터를 변수 `x`로 받습니다.
 - 두 번째 전달인자는 표본의 개수로 기본값 3이 할당되어 있어 값을 전달하지 않더라도 함수 내부에서 `s.size` 변수에 3이 저장됩니다.
 - 만일 사용자가 입력한 값이 있을 경우 해당 값이 `s.size`에 저장됩니다.

예제 : 모비율에 대한 점추정을 위한 표본비율의 분포

- 참(TRUE)과 거짓(FALSE)으로 구성된 벡터에 대해 수치연산을 실시하면,
 - TRUE는 1로, FALSE는 0으로 처리됩니다.
 - 이로 인해 sum() 함수와 결합할 경우 TRUE의 개수를 구할 수 있습니다.
 - p.even은 이를 이용하여 3개의 표본 중 짝수의 비율을 구합니다.

```
15.phat <- apply(smps.all, 1, p.even)
```

- **Step #3)** 출현 가능한 표본에서 짝수의 비율을 구해 변수 phat에 저장합니다
 - 15줄 : apply() 함수를 적용하여 각 행별로 p.even 함수를 적용하여 행별 짝수의 비율을 phat에 저장합니다.

예제 : 모비율에 대한 점추정을 위한 표본비율의 분포

▣ apply() 함수

- 기존 자료구조에 사용자가 지정한 함수를 적용하여 결과를 반환합니다.
- apply()에서 사용하는 자료는 다차원 자료를 사용합니다.
 - ▣ 행과 열로 구성된 표 형태의 자료는 2차원이며, 설명은 이를 중심으로 합니다.
- apply() 지정한 차원별로 함수를 적용합니다.

• R 도움말

`apply(X, MARGIN, FUN, ...)`

- X는 함수를 적용할 자료를 전달합니다.
- MARGIN : 행과 열로 구성된 경우 1은 행, 2는 열을 나타냅니다.
- FUN : 적용할 함수로 사용자 정의 함수도 사용할 수 있습니다.
 - ▣ 함수의 이름을 적습니다. 함수가 전달인자를 필요로 할 경우 뒤에 적습니다.
- 15줄에서 apply() 함수는 smps.all의 행별로(MARGIN=1) 합(FUN=sum)을 구합니다.

예제 : 모비율에 대한 점추정을 위한 표본비율의 분포

- **Step #4) 표본비율 \hat{p} 의 기대값과 분산을 구합니다.**

- 17줄 : phat의 평균, 즉 표본비율의 기대값을 구합니다.
- 18줄 : 모집단에서 짝수의 비율은 0.5입니다.
 - 표본비율의 기대값이 모비율과 동일합니다.

```
> mean(phat)
[1] 0.5
> ( p.p <- 0.5 )
[1] 0.5
```

- 19줄 : phat의 분산, 즉 표본분산을 구합니다.
- 20줄 : 알고있는 모비율을 이용하여 표본비율의 분산을 구합니다.
 - 표본비율의 분산과 $\frac{nP(1-P)}{n}$ 이 동일합니다.

```
> var.p(phat)
[1] 0.08333333
> ( p.p*(1-p.p)/3 )
[1] 0.08333333
```

예제 : 모비율에 대한 점추정을 위한 표본비율의 분포

- **Step #5) 표본비율 \hat{p} 의 표준오차를 구합니다.**
 - 21줄 : 표본분산에 제곱근을 구해 표준오차를 구합니다

```
> sqrt(var.p(phat))  
[1] 0.2886751
```



02. 구간추정

: 통계에 필요한 계산과 그림을 멋지게
처리하는 도구

1. 신뢰구간에 대해 학습한다.
2. 모집단의 분산에 대한 정보유무 에 따라 구하는 구간추정 방법에 대해 학습한다.

신뢰구간

- 점추정의 단점

- 점추정을 위해 사용하는 표본은 확률표본으로 추출된 표본에 따라 그 값이 달라집니다.
- 구간추정
 - 모수의 참값이 존재할 것으로 추정되는 구간을 표본으로부터 구하여 추정하는 방법

- 신뢰구간

- 구간추정을 위해 표본으로부터 구한 하한과 상한을 각각 $\widehat{\theta}_L, \widehat{\theta}_U$ 이라 할 때, $0 < \alpha < 1$ 인 α 에 대해

$$P(\widehat{\theta}_L < \alpha < \widehat{\theta}_U) = 1 - \alpha$$

를 만족하는 구간 $(\widehat{\theta}_L, \widehat{\theta}_U)$ 을 “모수 θ 에 대한 $100(1 - \alpha)\%$ 신뢰구간”이라 부르고, $(1 - \alpha)$ 를 신뢰수준이라 부릅니다.

구간추정 예

- 모집단의 분산을 알 때 모평균의 구간추정

- 평균이 μ 이고 분산이 σ^2 인 모집단으로 부터 추출한 n 개의 확률표본

X_1, X_2, \dots, X_n 의 표본평균의 분포 \bar{X} 는 평균이 μ 이고 분산이 $\frac{\sigma^2}{n}$ 인 정규분포를 따릅니다.

- 이를 표준정규분포로 변환하는 것으로 시작합니다.

- \bar{X} 에서 기대값인 μ 를 빼고 표준편차인 $\frac{\sigma}{\sqrt{n}}$ 로 나눈 표준화 변환을 사용합니다.

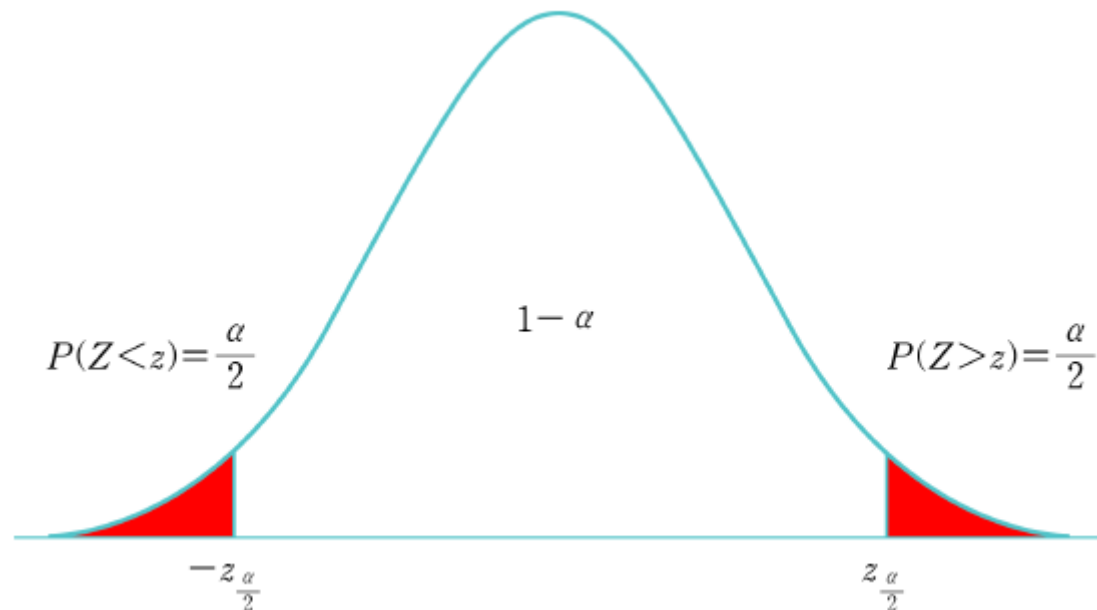
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1^2)$$

- 표준정규분포표를 이용하여 하한과 상한을 구한 후 원래의 정규분포로 돌아와 구할 것입니다.
 - 이제 이를 이용하여 구간추정에 대해 알아보시다.

구간추정 예

- 표본평균의 분포를 표준정규분포로 변환한 $100(1 - \alpha)\%$ 신뢰구간은 다음과 같이 하한으로 $\widehat{\theta}_L = -z_{\alpha/2}$, 상한으로 $\widehat{\theta}_U = z_{\alpha/2}$ 를 갖는 영역입니다.

$$P(\widehat{\theta}_L < Z < \widehat{\theta}_U) = P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$



구간추정 예

▣ 모평균에 대한 95% 신뢰구간

- 95% 신뢰구간 : 신뢰수준 $(1 - \alpha)$ 가 0.95인, 즉 α 를 0.05로 하는 신뢰구간
- 구하는 과정

① 하한과 상한

▣ $\widehat{\theta}_L = -z_{\alpha/2}$ 은 $P(Z < -z) = 0.025$ 가 되는 표준정규분포의 값으로 약 -1.96입니다.

▣ $\widehat{\theta}_U = z_{\alpha/2}$ 은 표준정규분포의 좌우대칭으로 1.96임을 알 수 있습니다.

② ①로 부터 다음임을 알 수 있습니다.

▣ $P(-z_{0.025} < Z < z_{0.025}) = P(-1.96 < Z < 1.96) = 0.95$

③ 원래의 정규분포에서 신뢰구간을 구하기 위해 다음을 계산합니다.

▣ $P(-z_{0.025} < Z < z_{0.025}) = P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$

구간추정 예

- ④ ③의 식을 전개하여 모평균에 대한 95% 신뢰구간을 구합니다.

$$\begin{aligned}
 0.95 &= P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) \\
 &= P\left(-1.96 \sigma/\sqrt{n} < \bar{X} - \mu < 1.96 \sigma/\sqrt{n}\right) \\
 &= P\left(-\bar{X} - 1.96 \sigma/\sqrt{n} < \mu < -\bar{X} + 1.96 \sigma/\sqrt{n}\right) \\
 &= P\left(\bar{X} + 1.96 \sigma/\sqrt{n} > \mu > \bar{X} - 1.96 \sigma/\sqrt{n}\right) \\
 &= P\left(\bar{X} - 1.96 \sigma/\sqrt{n} < \mu < \bar{X} + 1.96 \sigma/\sqrt{n}\right)
 \end{aligned}$$

- 이상으로부터 모평균에 대한 95% 신뢰구간은 다음과 같습니다.

$$\left(\bar{X} - 1.96 \sigma/\sqrt{n}, \quad \bar{X} + 1.96 \sigma/\sqrt{n}\right)$$

- 모평균에 대한 95% 신뢰구간은 하한으로 '(표본평균)-(1.96배의 표준오차)'를, 상한으로 '(표본평균)+(1.96배의 표준오차)'를 갖는 구간입니다

구간추정 예

예제 5-4 모평균에 대한 95% 신뢰구간

준비파일 | 05.95p.CI.R

- 실습내용

- 표준정규분포로부터 10개의 표본을 뽑아 95% 신뢰구간을 구하는 것을 100번 반복했을 때, 몇 개의 신뢰구간이 모평균 μ 을 포함할지 확인해봅니다.
- 신뢰구간의 의미를 생각해 봅시다.

```
1. set.seed(9)
2. n <- 10
3. x <- 1:100
4. y <- seq(-3, 3, by=0.01)
```

구간추정 예

- **Step #1) 필요한 변수들을 초기화합니다.**

- 1줄 : 동일한 난수를 생성하기 위해 난수의 초깃값을 9로 합니다.
- 2~4줄 : 표본의 크기를 10으로 하여 변수 n에 저장하고, x는 1부터 100까지의 표본추출 순서, y는 -3부터 3까지 0.01씩 증가하는 벡터로 저장합니다.

```
6. smps <- matrix(rnorm(n * length(x)), ncol=n)
```

- **Step #2) 표준정규분포로부터 난수를 생성합니다.**

- 6줄 : ① 표준정규분포로부터 '표본 개수(n=10) * 표본추출횟수(x의 크기=100)'인 1,000개의 난수를 생성합니다. (rnorm() 함수)
- ② 생성한 난수로 열의 개수가 10개인(ncol=n) 행렬을 만듭니다. (matrix() 함수)
 - ➔ 표준정규분포로부터 생성된 난수는 모두 1,000개이고, 이로부터 행이 100개이고 열이 10개인 행렬을 만듭니다. 이렇게 만들어진 행렬은 각 행별로 10개씩 추출한 표본의 역할을 할 것입니다.

구간추정 예

```

8. xbar <- apply(smps, 1, mean)
9. se <- 1 / sqrt(10)
10. alpha <- 0.05
11. z <- qnorm(1 - alpha/2)
12. ll <- xbar - z * se
13. ul <- xbar + z * se

```

- **Step #3) 각 표본추출로부터 평균, 하한 및 상한을 구합니다.**
 - 8줄 : 각 행별로 mean() 함수를 적용한 결과를 변수 xbar에 저장합니다.
 - 9줄 : 표준오차를 구합니다.
 - 분자는 모집단의 표준편차인 1, 분모는 표본의 개수인 10의 제곱근입니다.
 - 10줄 : 신뢰수준 95%는 α 를 0.05로 하므로 이를 변수 alpha에 저장합니다.
 - 11줄 : 하한의 z와 상한의 z 사이의 면적(확률)이 0.95가 되는 z 값을 구합니다.
 - 표준정규분포로부터 $P(Z < z) = 0.025$ 가 되는 $z_{\alpha/2}$ 를 구합니다.
 - 12, 13줄 : 하한과 상한으로 '표본평균 $\pm z_{\alpha/2}SE(\bar{X})$ '를 구합니다.

구간추정 예

```

15. plot( y, type="n", xlab="표본추출", ylab="z",
          xlim=c(1, 100), ylim=c(-1.5, 1.5), cex.lab=1.8 )
16. abline(h=0, col="red", lwd=2, lty=2)
17. l.c <- rep(NA, length(x))
18. l.c <- ifelse(l1 * u1 > 0, "red", "black")
19. arrows( 1:length(x), l1, 1:length(x), u1, code=3, angle=90,
            length=0.02, col=l.c, lwd=1.5 )

```

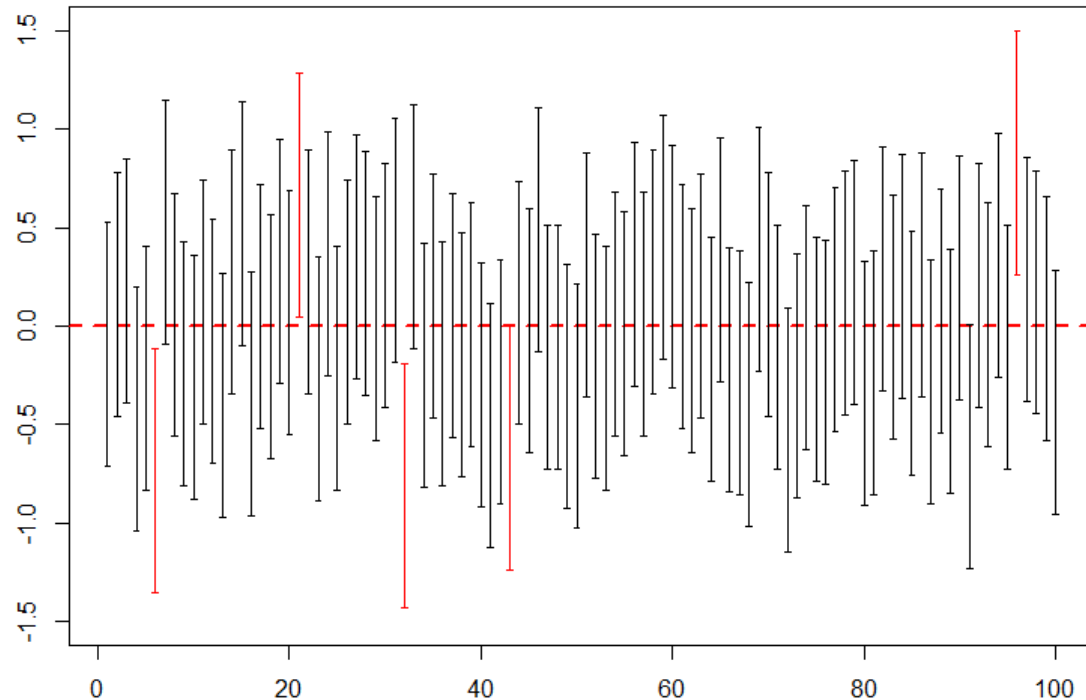
- **Step #4) 100번의 표본추출별로 신뢰구간을 그립니다.**

- 15줄 : 100개의 신뢰구간이 들어갈 빈 영역(type="n")을 그립니다.
 - x축에 각 시행별로 신뢰구간을 그립니다. 100회 시행 모두 표현해주기 위해 `xlim=c(0, 100)`으로 하였습니다.
 - y축은 -1.5부터 1.5까지(`ylim=c(-1.5, 1.5)`) 표준정규분포의 값이 들어가도록 하였습니다.
- 16줄 : 모집단의 평균인 0을 붉은색으로 그립니다.
 - `abline()` 함수에 `h=0`을 전달하면 y축의 값이 0인 수평선을 그립니다.

구간추정 예

- ▣ 17, 18줄 : 각 신뢰구간별 색깔을 지정합니다.
 - 각각의 신뢰구간이 평균을 포함하면 하한은 음수, 상한은 양수이므로 하한과 상한을 곱한 값이 음수가 됩니다. 이 경우 검은색으로 선을 그립니다.
 - 신뢰구간이 평균을 포함하지 않으면(하한과 상한이 모두 음수이거나, 양수이므로 두값이 곱이 양수이면) 빨간색으로 선을 그립니다.
 - 이를 신뢰구간별로 상황에 맞는 선의 색깔이 들어갈 벡터 변수 `l.c`를 생성합니다.
 - `ifelse()` 는 벡터 연산으로 조건식을 통해 생성되는 참과 거짓의 벡터의 크기와 동일한 크기를 받는 결과벡터를 반환합니다.(p. 253 참고)
 - `ifelse()` 함수는 첫번째 전달인자로 조건식을 전달하여 조건식이 참이면 두번째 전달인자의 값을, 거짓이면 세번째 전달인자의 값을 갖는 벡터를 생성합니다.
- ▣ 19줄 : 화살표의 머리 각을 90도(`angle=90`)로 하고 화살표의 시작점과 끝점에 머리가 생기도록(`code=3`) 하여 신뢰구간을 \cap 의 형태로 표현합니다

구간추정 예



• 신뢰구간의 의미

- ▣ 모집단에서 n 개의 확률 표본을 추출하는 것을 여러 번 실시하여 각각의 신뢰구간을 구하면 그들 중 약 95% 정도는 실제 모평균을 포함함을 의미합니다.
- ▣ 앞서 표본의 크기를 10으로 하여 신뢰구간을 구하는 것을 100번 반복했을 때 95번 정도는 실제 모평균을 포함하는 것을 보이고 있습니다.

구간추정 예

• 모집단의 분산을 모를 때 모평균의 구간추정

- 모집단이 미지의 평균과 분산을 갖는 정규분포를 따를 때 ,
이로부터 추출된 n 개의 확률표본 X_1, X_2, \dots, X_n 의 표본평균을 \bar{X} , 표본분산을 S^2 (표준편차 S) 이라 하면,
다음의 통계량 T 는 자유도가 $(n-1)$ 인 t 분포를 따름을 앞서 학습하였습니다.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

- 모집단의 분산을 모르는 경우 모집단의 표준편차 대신 표본의 표준편차를 사용하는 t -분포를 이용하여 모평균의 구간추정을 실시합니다.
- $100(1 - \alpha)\%$ 신뢰구간은 다음과 같이 하한으로 $\widehat{\theta}_L = -t_{\alpha/2, n-1}$, 상한으로 $\widehat{\theta}_U = t_{\alpha/2, n-1}$ 을 갖는 영역입니다.

$$P(\widehat{\theta}_L < T < \widehat{\theta}_U) = P(-t_{\alpha/2, n-1} < T < t_{\alpha/2, n-1}) = 1 - \alpha$$

- 모집단의 분산을 모르고 표본의 크기가 5일 때 모평균에 대한 95% 신뢰구간을 구해봅시다.

구간추정 예

▣ 구하는 과정

① 하한과 상한

- ▣ $\widehat{\theta}_L = -t_{0.025, 4}$ 은 $P(Z < -t(4)) = 0.025$ 가 되는 t-분포의 값으로 약 -2.78입니다.
- ▣ $\widehat{\theta}_U = t_{0.025, 4}$ 은 t-분포의 좌우대칭성을 이용하여 약 2.78입니다.

② ①로 부터 다음임을 알 수 있습니다.

- ▣ $P(-t_{0.025, 4} < T < t_{0.025, 4}) = P(-2.78 < T < 2.78) = 0.95$

③ 신뢰구간을 구하기 위해 ②의 식에서 $T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{S/\sqrt{4}}$ 로 치환합니다.

$$P(-t_{0.025, 4} < T < t_{0.025, 4}) = P\left(-2.78 < \frac{\bar{X} - \mu}{S/\sqrt{4}} < 2.78\right) = 0.95$$

구간추정 예

- ④ ③의 식을 전개하여 모평균에 대한 95% 신뢰구간을 구합니다.

$$\begin{aligned}
 0.95 &= P\left(-2.78 < \frac{\bar{X} - \mu}{S/\sqrt{4}} < 2.78\right) \\
 &= P\left(-2.78 \frac{S}{\sqrt{4}} < \bar{X} - \mu < 2.78 \frac{S}{\sqrt{4}}\right) \\
 &= P\left(-\bar{X} - 2.78 \frac{S}{\sqrt{4}} < \mu < -\bar{X} + 2.78 \frac{S}{\sqrt{4}}\right) \\
 &= P\left(\bar{X} + 2.78 \frac{S}{\sqrt{4}} > \mu > \bar{X} - 2.78 \frac{S}{\sqrt{4}}\right) \\
 &= P\left(\bar{X} - 2.78 \frac{S}{\sqrt{4}=2} < \mu < \bar{X} + 2.78 \frac{S}{\sqrt{4}=2}\right)
 \end{aligned}$$

- 이를 일반화하여 모분산을 모를 경우 모평균에 대한 95% 신뢰구간은 다음과 같습니다.

$$\left(\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}, \quad \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}\right)$$

구간추정 예

예제 5-5 모평균에 대한 95% 신뢰구간(모분산을 모를 때)

준비파일 | 06.Cl.t.R

• 실습내용

- 만 7세의 어린이 중 부모의 동의를 얻은 학생 중에서 10명을 표본으로 추출하여 머리 둘레를 측정한 결과는 다음과 같습니다.

520	498	481	512	515
542	520	518	527	526

- 이 자료로부터 모평균에 대한 95% 신뢰구간을 구해 봅시다.
 - 이를 위해 t-분포를 이용하여 신뢰구간을 구하는 사용자 정의 함수를 작성합니다.
 - 만들고자 하는 함수는 두 개의 전달인자를 받습니다.
 - 첫번째 전달인자 : 표본으로부터 관찰된 자료들의 벡터
 - 두번째 전달인자(alpha) : 신뢰수준을 구하기 위한 α 를 전달받습니다. 만일 사용자가 이 값을 사용하지 않을 경우에는 0.05를 사용합니다(기본 전달인자).

구간추정 예

```
1. ci.t <- function(x, alpha=0.05) {  
2.   n <- length(smp)  
3.   m <- mean(x)  
4.   s <- sd(x)  
5.   t <- qt(1-(alpha/2), df=n-1)  
6.   ll <- m - t * (s / sqrt(n))  
7.   ul <- m + t * (s / sqrt(n))  
8.   ci <- c(1-alpha, ll, m, ul)  
9.   names(ci) <- c( "Confidence Level", "Lower limit",  
                    "Mean", "Upper limit" )  
10.  return( ci )  
11. }
```

구간추정 예

- **Step #1) t-분포를 이용하여 신뢰구간을 만드는 함수를 작성합니다.**
 - 1줄 : 함수의 이름은 ci.t입니다.
 - 신뢰구간을 구할 자료들이 들어있는 벡터를 전달받아 변수 x에 저장하여 사용하고, 오류의 확률 α 는 기본전달인자로 0.05를 기본값으로 합니다.
 - 2줄 : 표본 크기를 변수 n에 저장합니다.
 - 3줄 : 표본의 평균을 변수 m에 저장합니다.
 - 4줄 : 표본의 표준편차를 변수 s에 저장합니다.
 - 5줄 : $t_{\alpha/2, n-1}$ 의 값을 구하기 위해 qt() 함수를 사용합니다.
 - 예제에서는 $1-0.05=0.975$ 가 되는 t값을 구하고($t_{0.975, n-1}$) 변수 t에 저장합니다.
 - t-분포의 좌우대칭을 이용하여 변수 t의 음수값 -t는 $t_{0.025, n-1}$ 이 됩니다.
 - 6줄 : $\bar{X} - t_{\alpha/2, n-1} s/\sqrt{n}$ 를 구해 변수 ll에 저장합니다(신뢰구간의 하한).
 - 7줄 : $\bar{X} + t_{\alpha/2, n-1} s/\sqrt{n}$ 를 구해 변수 ul에 저장합니다(신뢰구간의 상한).

구간추정 예

- 8줄 : 앞서 구한 신뢰수준, 하한, 평균, 상한으로 구성된 벡터 ci를 생성합니다.
- 9줄 : 벡터 ci의 각 값의 이름을 주어 출력 시 알아보기 쉽게 합니다.
 - names() 함수를 이용하여 이름이 있는 벡터를 생성합니다.
- 10줄 : 함수를 호출한 상대방에게 벡터 ci를 반환해줍니다.

```
13. smp <- c(520, 498, 481, 512, 515, 542, 520, 518, 527, 526)
14. ci.t(smp)
15. ci.t(smp, 0.1)
```

• Step #2) 앞서 만든 함수를 사용해 봅시다.

- 13줄 : 표본으로부터 관찰한 자료들을 벡터로 생성하고, 변수 smp에 저장합니다.
- 14줄 : ci.t 함수에 위의 자료 smp를 전달하고, 함수 수행의 결과를 보여줍니다.
 - alpha에 기본전달인자인 0.05 사용하여 95% 신뢰구간을 구합니다.
- 15줄 : 90% 신뢰구간을 구하기 위해 alpha 값을 0.1로 변경합니다.
 - 95% 신뢰구간보다 그 폭이 줄어드는 것을 확인할 수 있습니다

구간추정 예

```
> ci.t(smp)
```

Confidence Level	Lower limit	Mean	Upper limit
0.95	503.98	515.90	527.82

```
> ci.t(smp, 0.1)
```

Confidence Level	Lower limit	Mean	Upper limit
0.9000	506.2408	515.9000	525.5592



6장을 위한 준비

: 외부로부터 자료 가져오기 및 표본추출

외부로부터 자료가져오기

- 외부에서 제공하는 데이터 파일을 R에서 불러옵니다.

- 이번에는 엑셀로 저장된 파일을 앞서 통계청의 마이크로 데이터 통합서비스 자료로 부터 다운받은 csv 파일의 형태로 저장하여 R에서 읽어봅시다.
- 사용할 데이터는
다양한 국가 표준 · 인증 · 제품안전정보 · 기술규제 관련 정책 등을 담당하는 국가기술 표준원에서는 한국인 인체표준 정보를 DB화하고, 한국인이 쓰기에 편리한 제품개발과 생활공간 디자인에 인체표준정보를 제공하기 위한 '**한국인 인체치수조사보급사업**'을 실시하며, 측정데이터를 '**한국인 인체치수조사 sizekorea**' 웹 (<http://sizekorea.kr>) 을 통해 공개하고 있는 데이터 입니다.
- 최근 7차 인체치수조사 데이터가 등록되었으나 지난 6차 자료를 사용할 것입니다.
 - 과거의 데이터를 제공하는 것도 의미있는 일입니다. 이를 통해 한국인의 인체치수가 어떻게 변했는지 확인해 볼 수도 있으니까요....

외부로부터 자료가져오기

예제 5-6 sizekorea 데이터 가져오기

• 실습내용 및 필요사항

- sizekorea의 데이터를 사용하여 우리가 필요로 하는 데이터를 추출해봅시다.
- 주소는 <http://sizekorea.kr> 으로 접속하면 먼저 “회원가입”을 통해 아이디를 생성하고 로그인 합시다.

한국인 인체치수조사

Size Korea

로그인 | **회원가입**

사이즈 코리아 | 측정데이터 검색 | 인체치수조사 보고서 | 표준인체 측정법 | 3차원인체형상 | 전자민원실

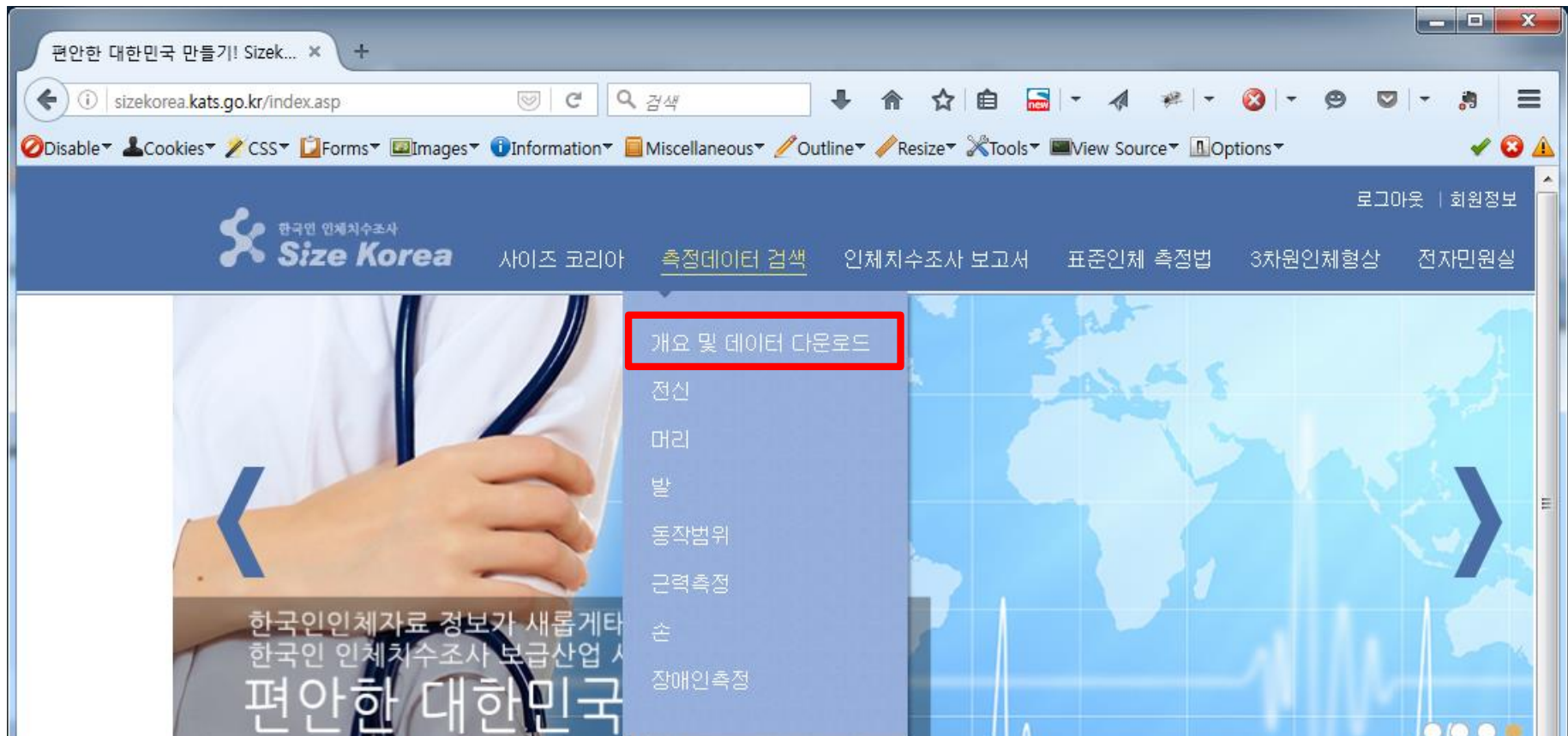
한국인인체자료 정보가 새롭게 태어납니다.
한국인 인체치수조사 보급산업 사이즈 코리아
편안한 대한민국을 위하여!

● ○ ○ ○ ○

외부로부터 자료가져오기

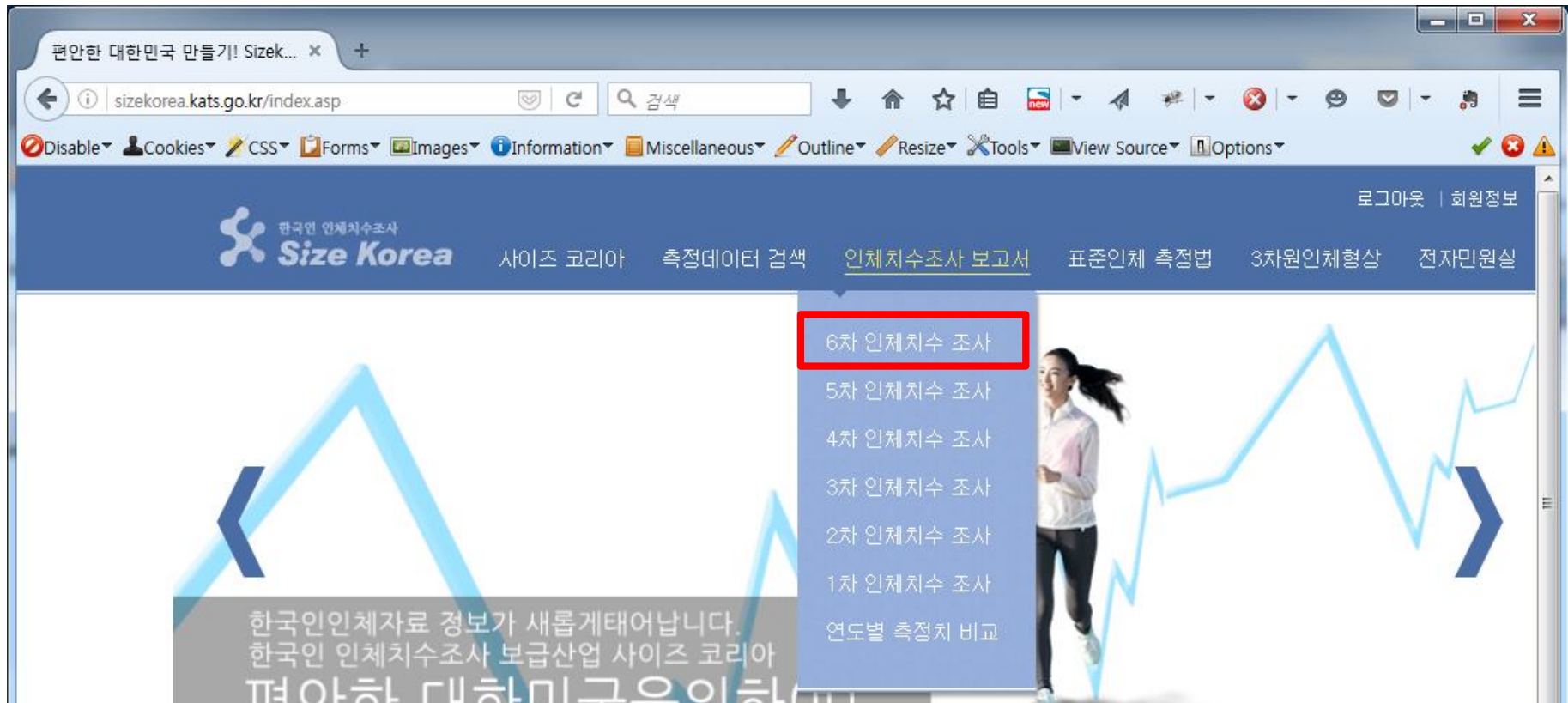
• Sizekorea

- 최신 측정데이터는 상단 메뉴의 '측정데이터 검색' ▶ '개요 및 데이터 다운로드'에서 다운로드 가능합니다



외부로부터 자료가져오기

- 우리가 사용할 데이터는 지난 “6차 인체치수 조사” 입니다.
 - ‘인체치수조사 보고서’ ▶ ‘6차 인체치수조사’를 선택합니다



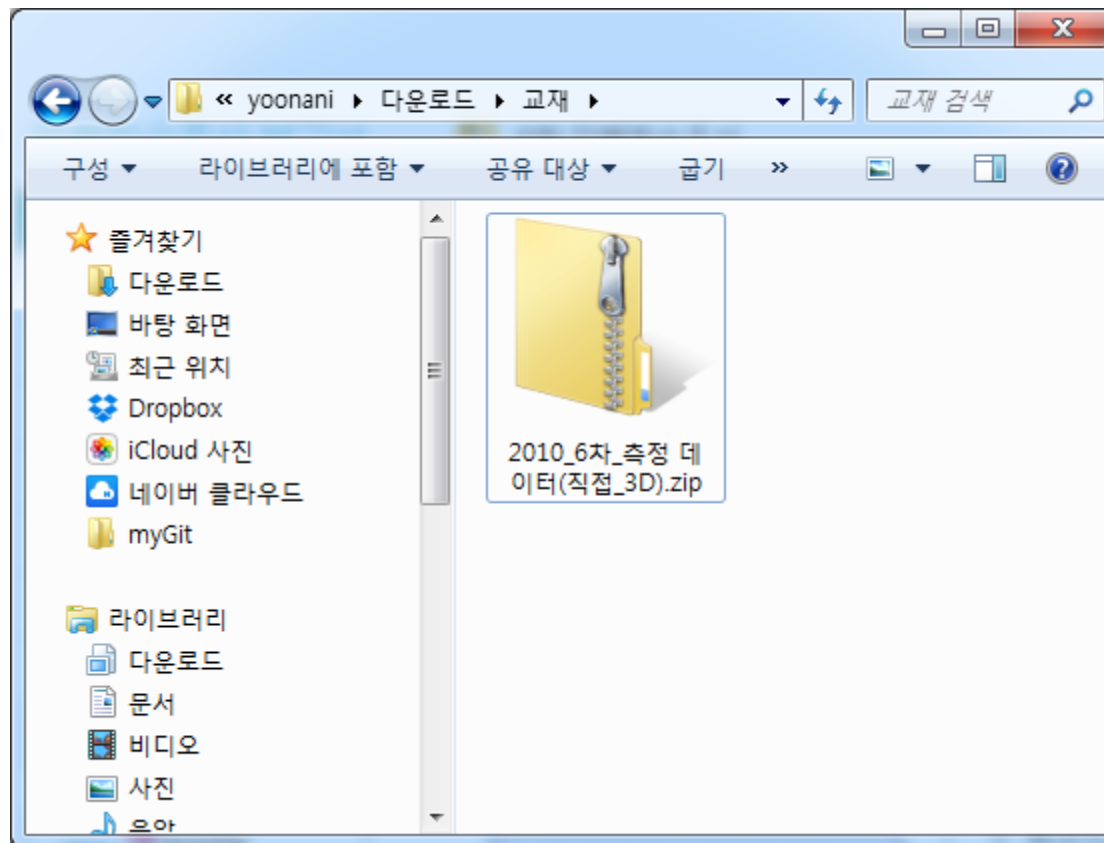
외부로부터 자료가져오기

'6차 인체치수조사' 하단의
'제6차인체수치데이터 다운로드'
를 클릭하여 자료를 받습니다

Size Korea 한국인 인체치수조사 사이트의 6차 인체치수조사 결과 페이지입니다. 페이지 상단에는 Size Korea 로고와 '사이즈 코리아', '측정데이터 검색' 등이 표시되어 있습니다. 중앙에는 '인체치수조사 보고서'라는 제목이 있으며, 그 아래에는 '6차 인체치수조사'를 포함한 목록이 있습니다. 목록에는 '5차 인체치수조사', '4차 인체치수조사', '3차 인체치수조사', '2차 인체치수조사', '1차 인체치수조사', '년도별 측정치 비교' 등이 포함되어 있습니다. 목록 아래에는 '인체측정표준화용어 다운로드' 버튼과 '문의전화 043. 870.5625' 정보가 있습니다. 오른쪽에는 '6차 인체치수조사'라는 제목과 함께 한 사람이 스키를 타는 사진이 있습니다. 사진 아래에는 '제 6차 결과 보고서 다운로드'와 '제6차 인체치수데이터 다운로드' 버튼이 있습니다. '제6차 인체치수데이터 다운로드' 버튼은 빨간색 테두리로 강조되어 있습니다.

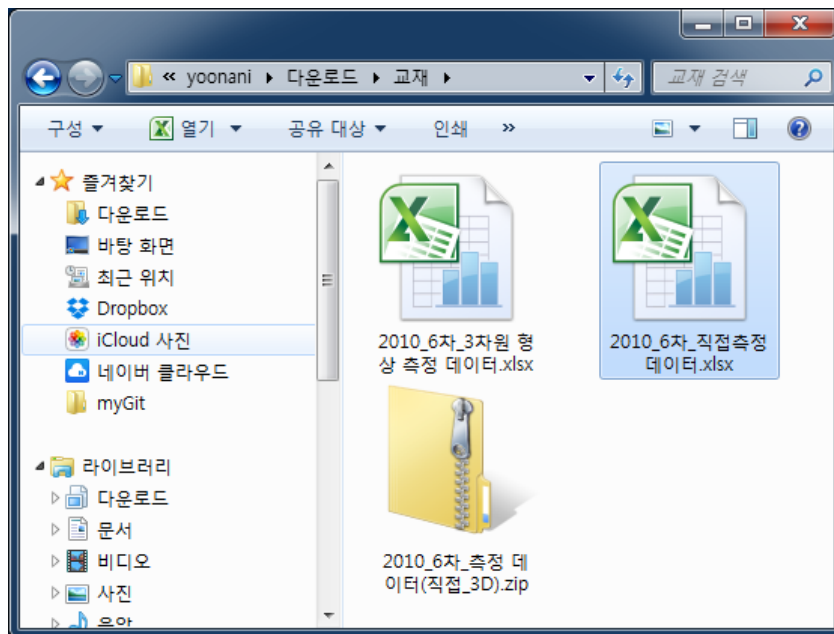
외부로부터 자료가져오기

- ▣ 다운로드 받은 파일은 압축파일(.zip)로 되어 있습니다
 - 압축을 해제합니다.



외부로부터 자료가져오기

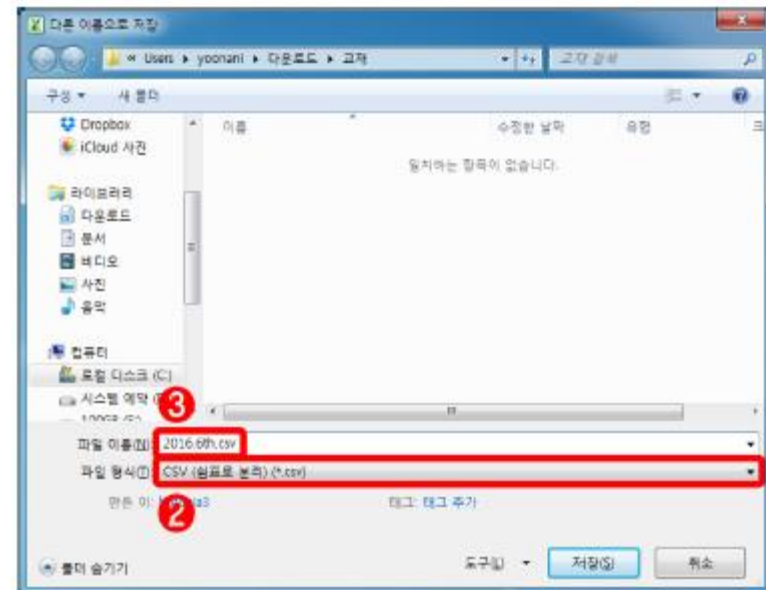
- 압축을 해제하면 '직접측정 데이터'와 '3차원 형상측정' 데이터 두 개의 엑셀파일(.xlsx)이 나옵니다.
 - '2010_6차_직접측정 데이터'를 열어봅시다.
 - 남성과 여성의 두 개의 sheet로 되어 있고, 전체 14,016명으로부터 139개 항목을 측정한 상당히 큰 데이터입니다



	A	B	C	D	E	F	G	H
1	성별	나이	101:오른쪽어깨경사각	102:왼쪽어깨경사각	103:머리위로뺨은주먹높이	104:키	105:눈높이	106:목뒤높이
2	남	23	22	24	2088	1740	1616	1475
3	남	22	24	18	2002	1722	1596	1439
4	남	24	26	26	2054	1788	1665	1521
5	남	23	28	27	2054	1770	1640	1495
6	남	23	27	27	2054	1697	1582	1425
7	남	24	26	20	2073	1781	1663	1512
8	남	23	20	24	1958	1673	1569	1416
9	남	20	22	22	2138	1830	1703	1551
10	남	23	15	12	2031	1710	1602	1462
11	남	26	24	23	2021	1726	1615	1450
12	남	25	15	16	2015	1704	1589	1464
13	남	22	17	18	2149	1778	1669	1503
14	남	24	20	17	2103	1784	1677	1545
15	남	24	16	16	2100	1786	1664	1508
16	남	21	18	14	2048	1757	1628	1494
17	남	23	21	18	2018	1754	1639	1474
18	남	28	28	25	2094	1796	1676	1509
19	남	23	20	18	1930	1669	1536	1397
20	남	26	26	21	2017	1730	1624	1475
21	남	24	30	28	2059	1763	1638	1493
22	남	21	20	15	1998	1703	1595	1438
23	남	19	25	24	2015	1742	1627	1471
24	남	19	20	20	2042	1720	1605	1449
25	남	19	21	20	2048	1755	1627	1480

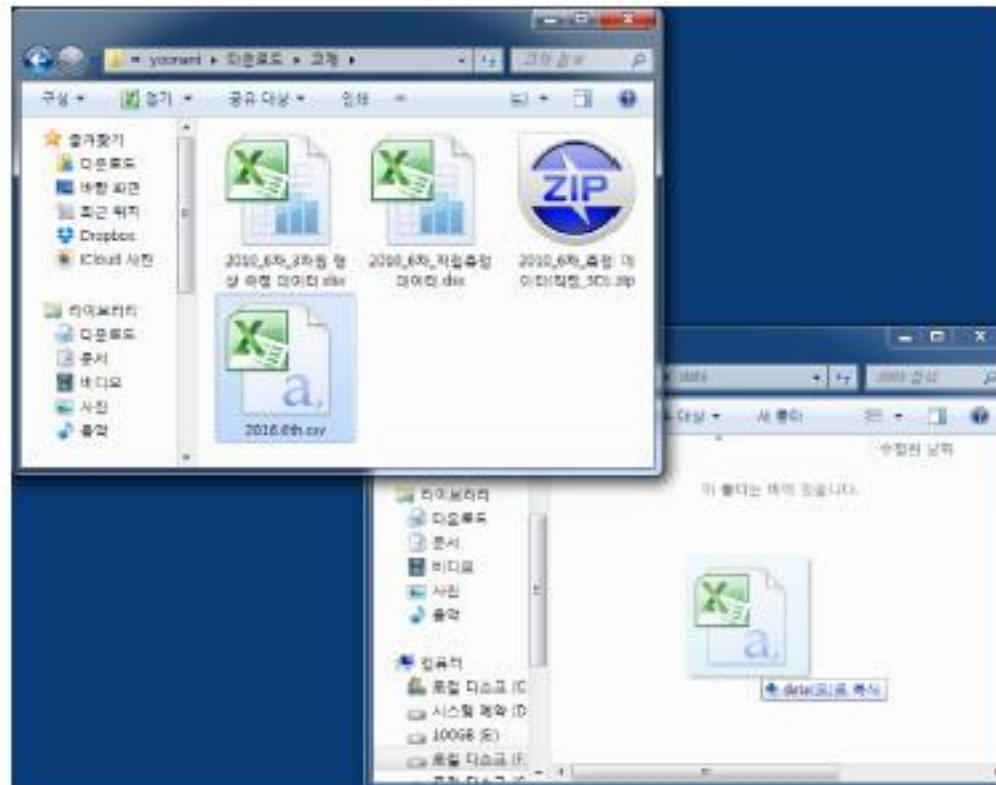
외부로부터 자료가져오기

- 파일에는 남성과 여성의 두 개의 탭이 있으며, 이들 중 ‘남성’ 시트에 있는 데이터를 R에서 읽어 들이기 위해 “남성” 시트에서
 - ① 엑셀에서 ‘다른 이름으로 저장’을 클릭한 후, 파일 저장창이 나오면
 - ② ‘파일 형식’을 ‘csv(쉼표로 분리)’로 선택합니다.
 - ③ ‘파일 이름’을 ‘2016.6th.csv’로 입력하고 저장합니다. (csv 파일로 저장시 현재 시트 저장여부와 엑셀의 각종 요소들이 저장되지 않음을 경고합니다.)



외부로부터 자료가져오기

- 변환한 데이터 파일을 data 폴더로 복사합니다.
 - 6장을 위한 프로젝트 생성 후 data 폴더에도 저장합니다.



R에서 불러오기와 추출하기

예제 5-7 R에서 불러오기와 추출하기

준비파일 | 07.retrieve.R

• 실습내용

- csv로 변환한 파일을 불러오고, 6장의 예제로 사용하기 위해 필요로 하는 자료들을 무작위로 선택합니다.

```
1. data <- read.csv("./data/2016.6th.csv", header=T)
2. str(data)
```

• Step #1) 앞서 sizekorea로부터 받은 파일을 불러옵니다.

- 1줄 : 앞서 저장한 csv 파일은 read.csv() 함수를 이용해 불러옵니다.
 - 자료의 첫 줄에 변수명이 있으므로 header=T를 주어 첫줄은 변수명으로 합니다.
 - header=T가 아닐 경우 1줄부터 데이터로 읽어 오고 변수명은 R이 V1, V2, V3, ... 의 형태로 임의로 만듭니다.
 - 만일 csv 파일에서 예를 들어 처음 5줄은 자료 설명이 있고 6줄 부터 자료가 시작될 경우 skip=5를 read.csv()에 넣어 처음 5줄을 건너 뛸 수 있습니다.

R에서 불러오기와 추출하기

- 2줄 : 불러온 데이터의 구조를 확인합니다.
 - 7,532개의 관찰대상으로부터 156개의 변수가 저장된 자료입니다.
 - 변수명이 숫자로 시작된 경우
 - read.csv() 함수가 변수명 앞에 대문자 X를 붙이고,
 - 변수명에 특수문자(이 경우 콜론 ':')나 띄어쓰기가 있는 경우 점('.')으로 변환합니다.
 - '101:오른쪽어깨경사각' → 'X101.오른쪽어깨경사각'
 - 변수가 156개나 되는 관계로 전부 출력되지 않을 수 있습니다

```
> str(data)
'data.frame':   7532 obs. of  156 variables:
 $ 성별           : Factor w/ 1 level "남": 1 1 1 1 1 1 1 1 1 1 ...
 $ 나이           : num  23 22 24 23 23 24 23 20 23 26 ...
 $ X101.오른쪽어깨경사각 : num  22 24 26 28 27 26 20 22 15 24 ...
 ...
 $ X324.벽면몸통두께    : num  250 246 214 234 264 276 263 248 308 295 ...
 $ X325.벽면어깨수평길이 : num  100 82 64 98 67 75 73 65 55 87 ...
 [list output truncated]
```

R에서 불러오기와 추출하기

```
4. tmp <- subset(data, 나이==7 )
5. height.p <- tmp$X104.키
```

- **Step #2)** 전체 데이터에서 필요한 부분만 추출합니다.
 - 우리가 사용할 자료는 7세 어린이들의 키 자료입니다.
 - 4줄 : subset() 함수를 이용해 조건에 맞는 행들만으로 구성된 데이터 프레임을 생성합니다.
 - subset 함수의 첫 번째 전달인자에는 원본 데이터 프레임의 이름이 들어가고,
 - 두 번째 전달인자에는 행 선택 조건이 들어갑니다. (나이 열의 값이 7인 자료 선택)
 - 7살아이들의 자료만 추출하기 위해 데이터 프레임 data의 '나이' 열의 값이 7인 행들을 선택(나이==7)하여 tmp라는 이름의 데이터 프레임으로 저장합니다.
 - 교재에서는 나이변수가 데이터프레임 data의 열임을 강조하기 위해 data\$나이로 표기하였으나, subset() 함수에서는 슬라이드처럼 사용합니다.
 - 5줄 : 4줄에서 작성한 7세 어린이의 여러 측정변수 중 '키' 변수를 선택하여 height.p에 저장합니다.

R에서 불러오기와 추출하기

```
7. set.seed(9)
8. height <- height.p[sample(length(height.p), 15)]
9. height
```

- **Step #3)** 주어진 자료에서 15개의 확률 표본을 생성합니다.
 - 7줄 : 난수생성의 초깃값을 9로 합니다.
 - 8줄 : sample() 함수를 이용해 15개의 확률표본을 생성하고, 이를 height에 저장합니다.
 - sample() 함수의 첫 번째 전달인자에 특정 숫자를 전달하면, 숫자의 순서를 무작위로 배치합니다.
 - sample() 함수의 첫번째 전달인자는 벡터가 전달되며 위와 같이 단일 값을 넣는 경우 1:n의 벡터를 무작위로 섞습니다.
 - 두 번째 전달인자로 특정숫자를 전달하면 임의로 배치된 숫자들 중 앞에서 원하는 개수만큼 가져옵니다
 - 앞서 추출한 height.p에서 15개의 확률표본을 추출하는 경우와 같습니다.

R에서 불러오기와 추출하기

- 9줄 : 8줄에서 생성한 height.p에서 추출한 15개의 표본을 출력합니다

```
> height  
[1] 1196 1340 1232 1184 1295 1247 1201 1182  
1192 1287 1159 1160  
[13] 1243 1264 1276
```




Q & A



수고하셨습니다.