강의교안 이용 안내

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

한빛아카데미 Hanbit Academy, Inc.

CHAPTER 06

가설검정

한빛아카데미 Hanbit Academy, Inc.

Contents

6.1 가설검정

- 가설검정
- 판정의 기준
- 기각역 수립 방법

6.2 단일 모집단의 가설검정

- 단일 모집단의 평균에 대한 가설검정
- 단일 모집단의 모비율에 대한 가설검정

7장을 위한 준비

01. 가설검정

: 모수에 대한 가설 결정하기

1. 가설검정에 대해 학습한다.
2. 어떤 가설을 선택할 것인지 판정의기 준에 대해 이해한다.
3. 대안가설에 따른 기각역 수립 방법에 대해 학습한다.

가설검정

- 모수의 상태에 대한 여러 주장들 중 어떤 주장을 사실로 받아들일지를 결정하는 과정
- **예) '만7세 남자 어린이의 키의 평균이 1220mm'라는 기존에 알려진 모수의 상태를 현재에도 받아들일 수 있는지 알아봅시다.**
  - 모수의 참값을 구하는 것이 아니라 모수의 상태가 '만7세 남자 어린이의 키의 평균이 1220mm'라는 기존에 알려진 사실이 현재에도 유지되고 있는지 확인하고자 합니다.
  - 이를 위해 다음과 같이 만 7세 남자 어린이들을 모집단으로 한 15명의 어린이를 표본으로 추출하고 키를 조사하였습니다.

1196	1340	1232	1184	1295
1247	1201	1182	1192	1287
1159	1160	1243	1264	1276

가설검정

- **가설검정의 과정**
  - 모집단 특성의 상태에 대한 주장인 가설에 대해 표본으로부터 얻은 정보를 바탕으로 이를 채택할지 기각할지를 판단함으로써 모집단의 상태에 대해 결정하는 과정으로, 다음의 4단계를 거쳐 이뤄집니다.
    - 1단계 : 가설 수립
    - 2단계 : 표본으로부터 검정을 위한 통계량 계산
    - 3단계 : 가설 선택의 기준 수립
    - 4단계 : 판정

가설수립

- **가설검정에서 사용하는 가설의 종류**
  - 영가설(귀무가설, H_0)
    - 주로 기존에 알려진 것과 차이가 없음을 나타냅니다.
  - 대안가설(대립가설, H_1)
    - 주로 기존에 알려진 것과 차이가 있음을 나타냅니다
    - 연구자가 밝히고자 하는 가설로 연구가설이라고도 합니다.
  - 영가설과 대안가설 수립의 예

가설	내용	수식 표현
영가설 H_0	(만7세 남자 어린이의) 키의 평균은 1220mm이다.	$\mu_{키} = 1220(\mathrm{mm})$
대안가설 H_1	(만7세 남자 어린이의) 키의 평균은 1220mm가 아니다.	$\mu_{키} \neq 1220(\mathrm{mm})$

표본으로부터 검정을 위한 통계량 계산

- **검정통계량**
  - 영가설의 채택 및 기각 여부를 확인하기 위해 표본을 통해 관찰된 값을 사용하는 통계량입니다.
  - 검정통계량의 계산은 표본으로부터 관찰된 특성이며, 모수의 상태로 '영가설이 참'이라는 가정 하에 계산하고, 판정단계에서 이 가정을 유지할 것인지의 여부를 결정합니다.
    - '영가설이 참'이라는 가정을 받아들일 수 없을 때 영가설을 기각합니다.
    - '영가설이 참'이라는 가정을 받아들일 때는 영가설을 채택합니다.

표본으로부터 검정을 위한 통계량 계산

- 검정통계량 계산의 예 : 만7세 남자 어린이의 평균 키에 대한 가설검정
  - 모집단은 '만7세 남자 어린이의 키' 하나이고, 평균에 대한 가설검정을 하는 경우입니다(모집단이 한 개일 경우의 평균 검정).
  - 한 개의 모집단 특성의 평균에 대한 검정에서는 **모집단의 분산을 모를 때** 4장에서 학습한 t-통계량을 사용합니다.

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$$

  - 검정통계량 계산
    - $\bar{X}$는 표본평균, s는 표본표준편차, n은 표본의 개수로 표본으로부터 관찰합니다.
    - μ_0는 우리가 알고자 하는 모평균으로, 영가설로부터 1220mm를 가져옵니다.
      - 이 과정이 바로 '영가설이 참'이라는 가정을 의미합니다.
    - 표본으로부터 구한 검정통계량은 '영가설이 참'일 때 자유도가 n-1인 t-분포에서 관찰된 값입니다.

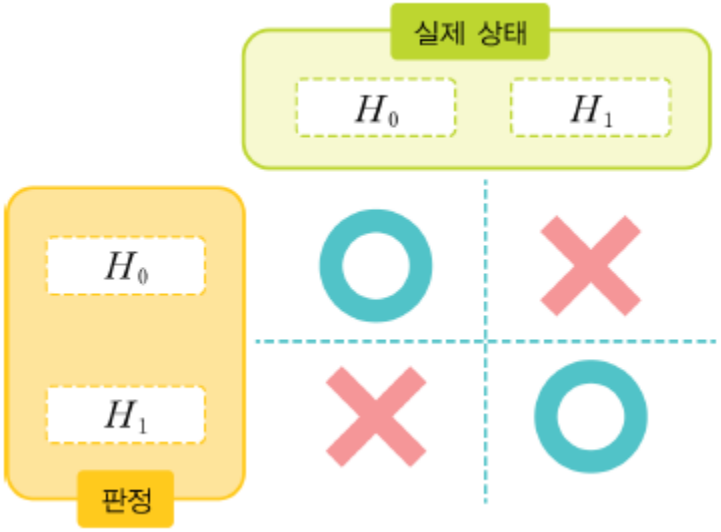
표본으로부터 검정을 위한 통계량 계산

- 관찰된 자료의 평균은 1230.533(mm)이고, 영가설 하에서 모평균은 1220(mm), 표본의 표준편차는 54.186(mm), 표본의 크기는 15으로 자유도가 14인 t-분포에서 약 0.727입니다.

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{1230.533 - 1220}{54.186/\sqrt{15}} \cong 0.727$$

가설 선택의 기준 수립 : 유의수준과 기각역

- **가설검정시 (판정의) 오류**
  - 실제 영가설이 참일 때 가설검정을 통해 대안가설을 선택하거나,
    실제 영가설이 거짓일 때 가설검정을 통해 영가설을 선택하는 경우



  - 제 1종 오류 : 영가설이 참인데 대안가설을 선택하는 오류
  - 제 2종 오류 : 영가설이 거짓인데 영가설을 선택하는 오류

가설 선택의 기준 수립 : 유의수준과 기각역

- **어떤 오류를 관리할 것인가?**
  - 법정에서는 국민참여재판을 신청한 피고인 A 씨에 대한 법정공방이 벌어지고 있습니다.
  - 변호인은 피고인이 무죄라고 주장하면서, 판사와 배심원단을 상대로 설전을 벌이고 있습니다.
  - 각종 증거와 반론이 오고간 법정공방을 마치고, 배심원단은 숙고 끝에 평결을 판사에게 전달하고 판사는 배심원단의 의견을 참고하여 판결을 내리고자 합니다.
  - 판사는 A 씨의 범죄에 대해 무죄 혹은 유죄를 판결함에 있어, 유죄인 상태를 영가설, 무죄인 상태를 대안가설이라고 하면 잘못된 판단은 다음의 두 가지가 있습니다.
    ① 실제 무죄이나 유죄 판결을 받아 양형에 따른 수감 생활(제2종 오류)
    ② 실제 유죄이지만 무죄 판결을 받아 사회로 돌아감(제1종 오류)

가설 선택의 기준 수립 : 유의수준과 기각역

- 여러 사람이 함께 어울리는 사회의 관점에서 보겠습니다.
  - ❶의 무죄이나 억울한 옥살이를 하게 되는 경우도 문제가 되겠지만,
  - 사회적인 관점에서만 보자면 ❷의 죄인이 죗값을 치루지 않고 유유히 법정을 나와 사회의 구성원이 되는 것이 더 큰 문제가 될 것입니다.
  - 제1종 오류의 경우, 영가설이 참이지만 참이 아니라고 주장하는 경우입니다.
    - 연구에서는 차이가 없으나 차이가 있다고 주장하는 경우로 제1종 오류가 더 심각한 상황이 될 것입니다.
    - 제 2종 오류의 확률 또한 중요한 내용을 담고 있지만, 우리의 범위를 벗어나므로 상위 과정에서 학습해 주세요. (좋은 검정법)
- 유의수준(α)
  - 제 1종 오류를 범할 확률의 최대 허용 한계를 유의수준이라고 합니다.
  - 연구에 따라 0.1, 0.05, 0.01 등 여러 기준이 있으나, 수업에서는 통상적으로 사용하는 유의수준인 0.05를 사용하겠습니다.

가설 선택의 기준 수립 : 유의수준과 기각역

- 유의수준의 역할 : 기각역 수립
  - 유의수준은 오류가 발생할 확률로써 이는 영가설 하에서 생성되는 표본분포에서의 확률을 나타냅니다.
  - 예 : 만7세 남자 어린이의 평균 키에 대한 가설검정에서의 유의수준을 정하고, 그 역할에 대해 살펴봅시다.
    - 대안가설
      - '(만7세 남자 어린이의) 키의 평균은 1220mm가 아니다($\mu_{키} \neq 1220$)'
      - 이 대안가설을 만족하는 상황은
        검정통계량 T가 1220mm보다 현저히 작은 경우($T < c_l$) 혹은 1220mm보다 현저히 큰 경우($T > c_u$) 에서 관찰된 두 가지입니다
    - '~보다 크다', '~보다 작다' 는 상대적인 개념으로 기준이 되는 값이 필요하고 유의수준이 그 기준을 제시해주는 역할을 합니다

가설 선택의 기준 수립 : 유의수준과 기각역

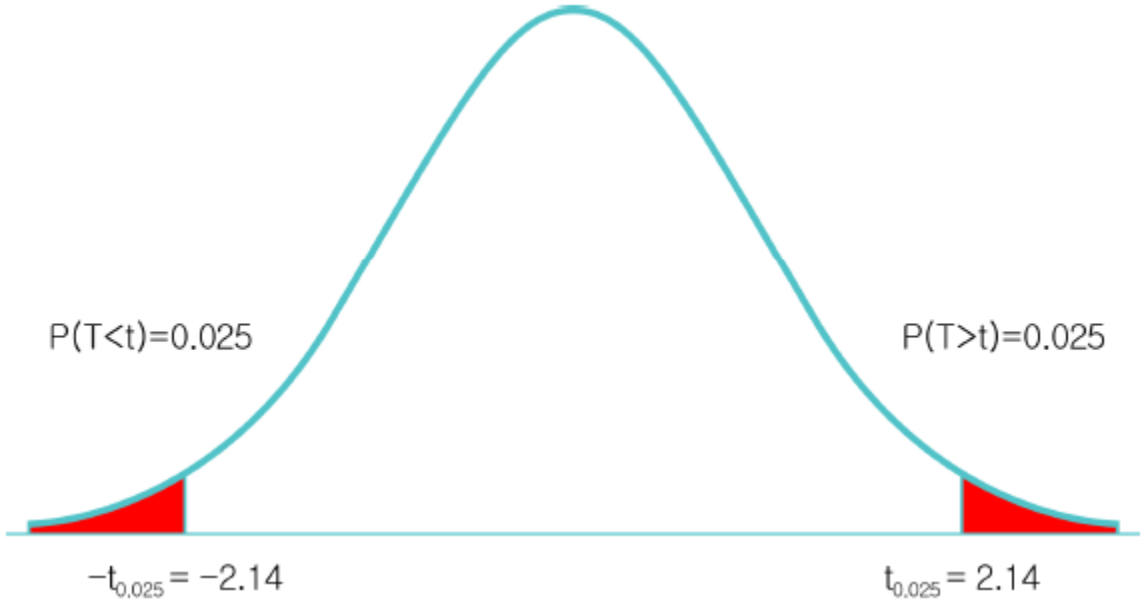
- 유의수준을 이용한 기준제시
  - 예에서는 대안가설에 의해 작은 쪽과 큰 쪽 두 곳의 기준이 필요합니다.
  ① 작은 쪽의 기준을 c_l이라 할 때, c_l은 영가설 하의 분포에서 $P(T < c_l) = {}^{\alpha}/_{2}$ 가 되게 하는 값입니다.
  ② 큰 쪽의 기준을 c_u이라 할 때, c_u은 영가설 하의 분포에서 $P(T > c_u) = {}^{\alpha}/_{2}$ 가 되게 하는 값입니다.
  - 이 기준에 따라 $\alpha = 0.05$라 했을 때 c_l 보다 작은 쪽의 확률이 0.025가 되게 하는 영가설 하에서 값을 R을 이용해 구해봅시다.

```
> qt(0.025, df=14)
[1] -2.144787
```

  - $P(T < c_l) = 0.025$인 자유도가 14인 t-분포에서의 값 c_l은 약 -2.14이며 $P(T > c_u) = 0.025$인 자유도가 14인 t-분포에서의 값 c_u는 t-분포의 좌우대칭을 이용하여 약 2.14임을 알 수 있습니다.

가설 선택의 기준 수립 : 유의수준과 기각역

- 임계값, 기각역과 채택역
  - 여기서 구한 두 값 $c_l = -2.14$와 $c_u = 2.14$ 를 임계값이라 합니다.
  - 분포의 중앙을 중심으로 임계값 바깥쪽의 영역 $T < c_l, T > c_u$를 **기각역**이라 합니다.
  - 분포에서 기각역이 아닌 영역 즉$c_l < T < c_u$을 **채택역**이라 합니다.
  - 다음의 그래프에서 붉은 영역이 기각역입니다.



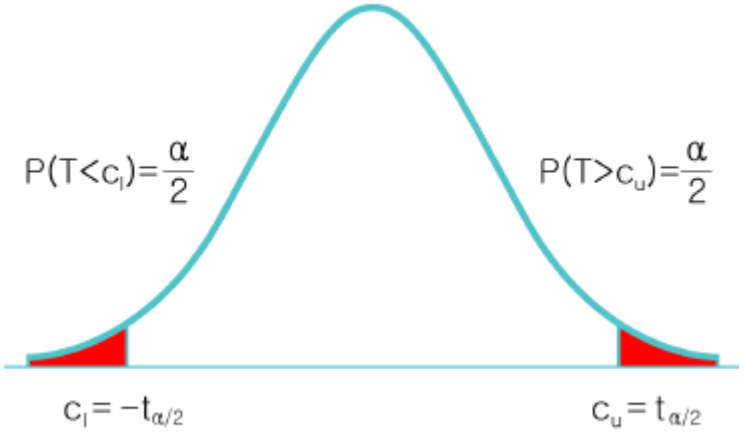
가설 선택의 기준 수립 : 유의수준과 기각역

- **기각역 설정**
  - 기각역은 대안가설에 따라 양쪽 혹은 한쪽에 생깁니다.
  - 양쪽에 기각역을 두고 검정하는것을 양쪽검정, 한쪽에 두고 검정하는 것을 한쪽검정이라고 합니다.
  - 모수 θ에 대한 가설검정에서 사용하는 가설은 다음 표와 같이 세 종류가 있으며, 각각에 해당하는 기각역을 정리해 보았습니다.(θ_0는 영가설하의 θ)

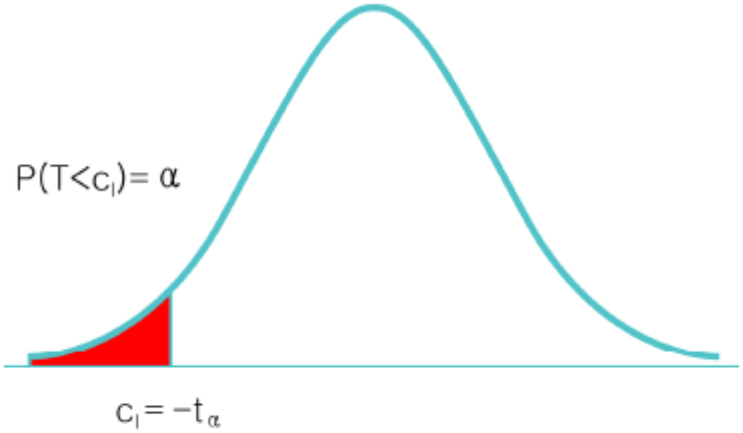
검정의 종류	영가설	대안가설	기각역과 유의수준
양쪽검정	$H_0 : \theta = \theta_0$	$H_1 : \theta \neq \theta_0$	$P(T > c_u) = \alpha/2$ $P(T < c_l) = \alpha/2$
(왼쪽) 한쪽검정	$H_0 : \theta \geq \theta_0$ $H_0 : \theta = \theta_0$	$H_1 : \theta < \theta_0$	$P(T < c_l) = \alpha$
(오른쪽) 한쪽검정	$H_0 : \theta \leq \theta_0$ $H_0 : \theta = \theta_0$	$H_1 : \theta > \theta_0$	$P(T > c_u) = \alpha$

가설 선택의 기준 수립 : 유의수준과 기각역

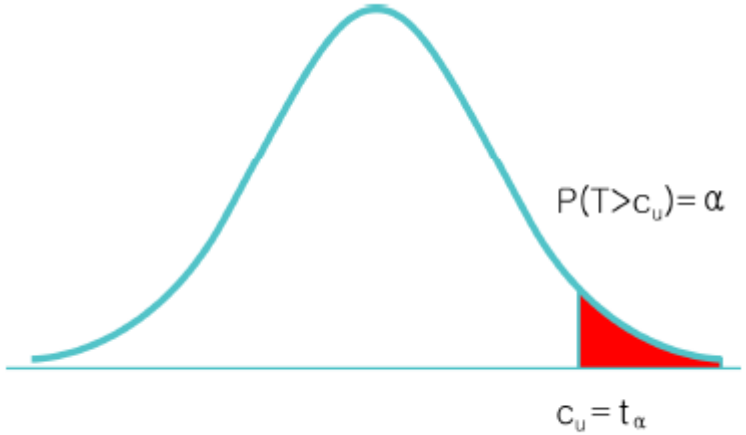
- t-분포 하에서 각각의 기각역은 다음과 같습니다.



(a) 양쪽검정 시의 기각역



(b) (왼쪽) 한쪽검정 시의 기각역



(c) (오른쪽) 한쪽검정 시의 기각역

판정

- **판정**
  - ① 앞서 **가설을 수립**하고
  - ② 가설을 검정하기 위해 표본을 추출하여 표본으로부터 **'영가설이 참'이라는 가정 하에 검정통계량**을 구하고,
  - ③ **유의수준**을 통해 기각역을 수립하여 판정이 기준이 되도록 하였습니다.
  - 이제 검정통계량과 기각역으로 영가설의 채택 여부를 판정해봅시다.

판정

- 검정통계량과 기각역을 이용한 판정
  - 가설
    - 영가설 : "만 7세 남자 어린이의 키의 평균은 1220mm이다."
    - 대안가설 : "만7세 남자 어린이의 키의 평균은 1220mm가 아니다." => 양쪽검정
  - 검정통계량 : 0.727 (자유도가 14인 t-분포에서)
  - 유의수준 : 0.05
    - 기각역은 $T < -2.14, T > 2.14$ 의 두 곳에 있습니다. (양쪽검정)
  - 판정
    ① 검정통계량이 기각역에 있으면, 영가설을 기각하고 대안가설 채택
    ② 검정통계량이 기각역에 있지 않으면, 영가설 채택(대안가설 기각)
    - 표본으로부터 구한 검정통계량은 0.753으로 기각역에 존재하지 않아 **영가설 채택**
      - 영가설이 참일 때($\mu_{키} = 1220\text{mm}$) 모평균에 대한 표본평균의 분포에서 충분히 발생할 수 있는 경우로 영가설이 참이라는 가정을 뒤집을 만한 근거가 되지 못합니다.

판정

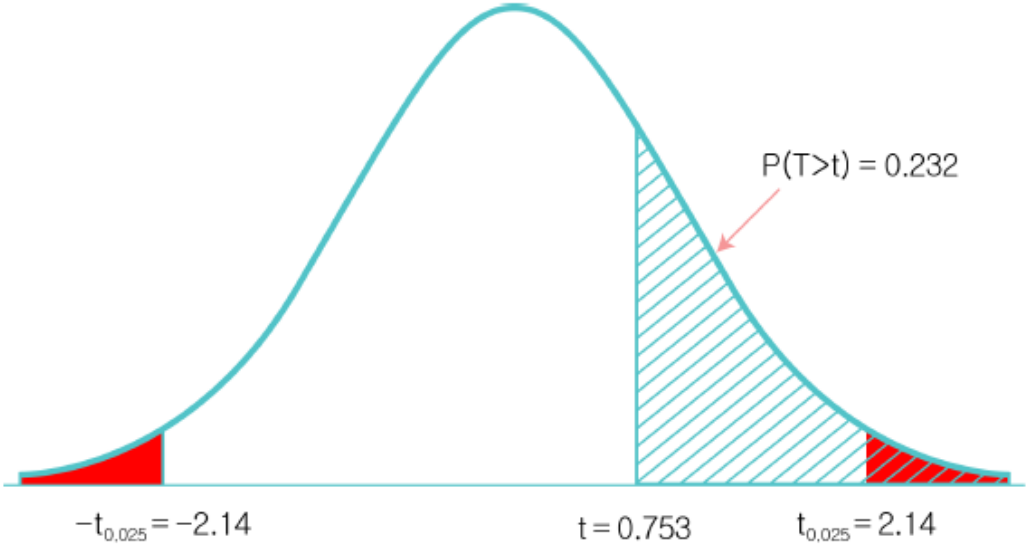
판정

- 유의확률과 유의수준을 이용한 판정방법
  - 영가설의 타당한 정도를 나타내는 확률에 대해 생각해봅시다.
    유의확률은 표본으로부터 계산된 검정통계량을 통해 구합니다.
  - 영가설을 기각할 수 있는 최소의 유의수준 역할로 영가설의 타당한 정도를 나타냅니다.
    - 유의확률이 크다면 영가설의 타당성이 높아 영가설을 채택하는 판단이 옳을 것이며,
    - 만일 그 값이 작다면 영가설의 타당성이 낮아 영가설을 기각하는 판단이 옳을 것입니다.
  - 유의확률 구하기
    - 양쪽검정에서는 검정통계량 t에 대해 $P(T > |t|)$
    - 한쪽검정에서 좌측 한쪽검정의 경우에는 $P(T < t)$ 로,
    - 한쪽검정에서 우측 한쪽검정의 경우에는 $P(T > t)$ 로 구합니다.

판정

- 앞서 구한 검정통계량에 대한 유의확률을 구해봅시다.
  - 자유도가 14인 t-분포에서 구한 검정통계량 0.753에 대한 유의확률은 $P(T > 0.753)$ 로 R에서 다음과 같이 구할 수 있습니다.

```
> 1 - pt(0.753, df=14)
[1] 0.2319624
```



판정

- 유의확률과 유의수준 비교 시 주의할 점
  - 양쪽검정을 실시했을 때 유의확률은 한쪽의 확률만 구한 것으로 유의확률을 유의수준의 반($\alpha/2$)과 비교하거나, 유의확률에 2배를 한 2×유의확률과 유의수준을 비교해야 합니다.
    - 앞선 예제의 경우 양쪽검정으로 검정통계량을 통해 구한 유의확률(약 0.24)과 우리가 정한 유의수준 0.05의 반인 0.025와 비교하여 판정을 내립니다.
      - 0.24는 0.025보다 크므로 **영가설을 채택**합니다.
    - 통계분석을 실시하는 통계 패키지(R, SAS, SPSS, Stata)들은 가설검정 시 연구자가 유의수준으로 어떤 값을 선택했는지 모르는 상태에서 계산을 하고 결과로 유의확률을 제시함으로써 연구자가 유의확률과 유의수준을 비교하여 결론을 내릴 수 있도록 합니다
      - SPSS의 경우 유의확률에 2배한 값을 계산해 줍니다.

결론 기술하기

- **결론 기술하기**
  - 가설검정에서는 판정 과정이 마지막이지만, 판정 결과를 다른 사람들에게 알리기 위해 판정의 근거와 함께 영가설의 채택 및 기각 여부를 가설검정의 결론으로 표현합니다.
  - 근거를 제시할 때는, 검정통계량 계산에 사용된 표본의 특성 및 계산된 검정통계량과 유의확률을 같이 제시하여 판정에 대한 근거로 활용합니다.
- **결론 기술 형태**
  ① 가설검정을 통해 밝히고자 하는 연구의 내용
  ② 표본으로부터 측정된 일반적 특성 및 검정통계량 계산의 근거가 되는 통계량
  ③ 검정통계량과 유의확률
  ④ 판정의 내용
  ⑤ 가설검정으로부터 알 수 있는 사실

결론 기술하기

- **결론 기술의 예 : 만 7세 어린이의 키의 평균에 대한 가설 검정**
  - ① "(만7세 남자 어린이의) 키의 평균이 1220mm"라는 기존 사실이 현재에도 유지되고 있는지 알아보기 위해
  - ② 15명의 7세 어린이를 표본으로 추출하여 키를 측정한 결과, 평균은 1230.53(mm), 표준편차는 54.186(mm)이었으며,
  - ③ 표본으로부터 구한 검정통계량은 0.753(유의확률 : 0.232)로 나타났습니다.
  - ④ 이는 유의수준 0.05에서 "(만7세 남자 어린이의) 키의 평균이 1220mm이다."라는 영가설을 기각할 수 없습니다.
  - ⑤ "(만7세 남자 어린이의) 키의 평균이 1220mm가 아니다."라는 대안가설에 대해 통계적으로 유의한 결론을 얻을 수 없었으며, (만7세 남자 어린이의) 키의 평균이 1220mm라는 기존의 사실은 여전히 유지되고 있는 것으로 판단됩니다.

02.단일 모집단의 가설검정

: 모평균과 모비율 검정

1. 단일 모집단의 평균에 대한 가설검정 방법에 대해 학습한다.
2. 단일 모집단의 비율에 대한 가설검정 방법에 대해 학습한다.

단일 모집단의 평균에 대한 가설검정

- **단일 모집단의 평균에 대한 가설 검정**
  - 하나의 모집단의 평균에 대한 검정 방법에 대해 알아봅시다.

예제 2 단일 모집단의 평균 검정 : 여아 신생아 몸무게의 평균 검정

- **내용**
  - 여아 신생아의 몸무게는 2800(g)으로 알려져 왔으나, 산모에 대한 관리가 더 세심해진 요즘 신생아의 몸무게가 증가할 것으로 판단되어, 이를 확인하고자 부모의 동의를 얻은 신생아 중 표본으로 18명을 대상으로 체중을 측정했습니다.
  - 즉, 연구자가 생각한 '여아 신생아의 체중이 2800(g)보다 크다'는 주장을 받아들일 수 있는지 검정해봅시다.

단일 모집단의 평균에 대한 가설검정

- 표본으로부터 측정된 여아 신생아의 몸무게는 다음과 같습니다.

3837	3334	2208	1745	2576	3208
3746	3523	3430	3480	3116	3428
2184	2383	3500	3866	3542	3278

- **가설검정**
  - 가설수립
    - 영가설 : (여아) 신생아의 체중은 2800g이다. (H_0 : $\mu_{몸무게} = 2800(g)$)
      - 기존의 알려진 사실인 여아 신생아의 몸무게 평균인 2800(g)을 영가설로 합니다.
    - 대안가설 : (여아) 신생아의 체중은 2800g보다 크다. (H_1 : $\mu_{몸무게} > 2800(g)$)
      - 연구자가 밝히고자 하는 것은 여아 신생아의 몸무게 평균이 기존보다 증가했다는 것으로 이를 대안가설로 수립합니다.

단일 모집단의 평균에 대한 가설검정

- 검정통계량(H_0 가 참이라는 가정 하에)
  - 모집단인 여아 신생아의 몸무게에 대해 분산에 대한 정보가 없습니다.
  - 이와 같이 모집단의 분산을 알지 못하고, 단일 모집단으로부터 추출된 단일 표본의 평균검정에서 사용하는 검정통계량은 자유도가 n-1인 t-분포를 따르는 T통계량

    ($T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$)을 사용합니다.
  - 검정통계량 계산을 위한 표본의 통계량
    - 표본의 개수 (n) : 18
    - 표본평균$(\bar{X})$: 3132.44(g)
    - 표본표준편차(s) : 631.5825(g)
  - 영가설 하의 모집단 평균 $\mu = 2800(g)$
  - $T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{3132.44-2800}{631.5825/\sqrt{18}} \cong 2.233$

단일 모집단의 평균에 대한 가설검정

- 유의수준 : $\alpha = 0.05$
  - 검정통계량이 따르는 분포는 자유도가 17인 t-분포이고,
  - 대안가설에 의해 '(오른쪽) 한쪽검정'으로 임계값은 $P(T > c_u) = 0.05$가 되는 c_u로, R을 통해 약 1.74 (유효숫자 소수점 셋째자리)임을 계산할 수 있습니다.

```
> c.u <- qt(1-alpha, df=n-1)
> c.u
[1] 1.739607
```

  - 검정통계량에 대한 유의확률($P(T > 2.233)$)은 다음과 같이 R을 통해 약 0.020 임을 알 수 있습니다.

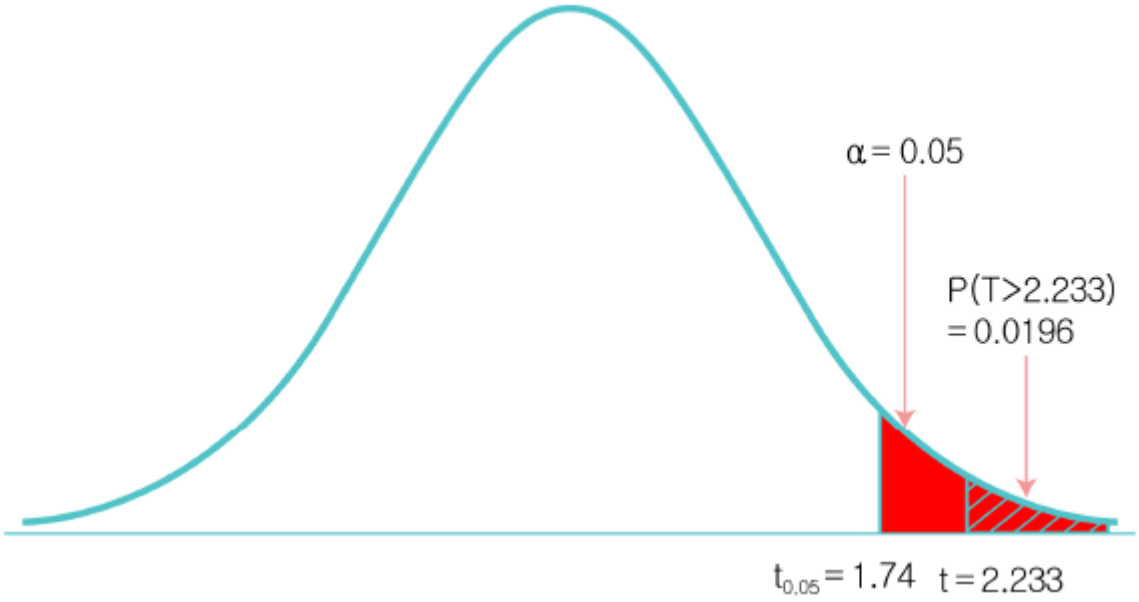
```
> 1-pt(2.233, df=n-1)
[1] 0.01964151
```

단일 모집단의 평균에 대한 가설검정

- 판정
  - 기각역을 이용한 판정
    - 앞서 구한 임계값은 1.74로, 검정통계량 2.233은 임계값보다 큰 기각역에 위치하여 영가설이 참이라는 가정을 받아들일 수 없습니다.
    - 영가설을 기각합니다.
  - 유의확률(p-value)을 이용한 판정
    - 앞서 구한 유의확률은 0.02로 유의수준 0.05보다 작습니다.
      이는 영가설을 사실로 받아들일 가능성이 낮은, 즉 영가설 하에서 발생하기 힘든 경우로 영가설이 참이라는 가정을 받아들일 수 없습니다.
    - 영가설을 기각합니다.

단일 모집단의 평균에 대한 가설검정

- 판정



단일 모집단의 평균에 대한 가설검정

- **결론**
  - 여아 신생아의 몸무게의 평균이 2800(g)보다 증가하였는지 알아보기 위해
  - 18명의 신생아로부터 측정한 몸무게의 평균과 표준편차는 3132.44±631.583 (g)으로 조사되었으며,
  - 이로부터 구한 검정통계량은 2.233(유의확률 0.02)으로 나타났습니다.
  - 따라서 "여아 신생아의 몸무게의 평균이 2800(g)보다 크다."는 통계적으로 유의한 결론을 얻을 수 있었습니다.
  - 이로부터 여아 신생아의 평균 체중은 기존에 알려진 2800(g)보다 증가한 것으로 여겨집니다.

단일 모집단의 평균에 대한 가설검정

예제 6-1 [예제 2]의 단일 모집단의 평균검정

준비파일 | 05.sample2.R

- 실습내용
  - 여아 신생아의 몸무게의 평균이 2800(g)보다 더 증가했다는 주장을 할 수 있는지 R 을 이용하여 검정하는 방법에 대해 알아봅시다.
  - 이를 위해
    - 먼저 R 코드를 이용하여 직접 가설검정에 필요한 각 값을 구해보고,
    - R이 제공하는 함수를 이용하여 검정의 결과를 얻고 각 출력물을 해석하는 방법에 대해 알아봅시다.

단일 모집단의 평균에 대한 가설검정

```
1. data <-
   read.table("http://www.amstat.org/publications/jse/datasets/bab
   yboom.dat.txt", header=F)
2. str( data )
3. names(data) <- c("time", "gender", "weight", "minutes")
4. tmp <- subset(data, gender==1)
5. weight <- tmp[[3]]
```

- **Step #1) 데이터 준비하기**
  - 1줄 : 열 구분자(열을 구분하는 기호)로 공백이나 탭을 사용할 경우 연속된 공백이나 탭을 하나의 구분자로 각 열을 구분하는 read.table() 함수를 이용합니다.
    - Journal of Statistics Education에서 제공하는 자료로 이를 가져와 변수 data에 데이터 프레임으로 저장합니다.
    - 해당 데이터 첫 줄부터 데이터가 시작(header=F)됩니다. (열 이름에 대한 정보가 없는 데이터 파일)

단일 모집단의 평균에 대한 가설검정

- 2줄 : 불러온 자료의 구조를 확인합니다
  - 44명의 관찰대상으로부터 4개의 변수가 관찰되었으며, 변수명에 대한 정보가 없어 변수의 이름을 R이 자동으로 V1, V2, V3, V4로 지정하였습니다.

```
> str( data )
'data.frame':  44 obs. of  4 variables:
 $ V1: int  5 104 118 155...
 $ V2: int  1 1 2 2 ...
 $ V3: int  3837 3334 3554 3838 ...
 $ V4: int  5 64 78 115 ...
```

- 3줄 : names() 함수를 이용하여 다음과 같이 변수의 이름을 지정해줍니다.

기존 변수명	V1	V2	V3	V4
사용할 변수명	times	gender	weight	minutes

  - names() 함수에 대해서는 이 장 뒷부분의 7장을 위한 준비를 참고해 주세요.

단일 모집단의 평균에 대한 가설검정

- 4줄 : 여아들만 선택하기
  - subset() 함수를 이용하여 변수 자료로 읽어온 데이터 프레임인 data의 gender 열의 값이 1인 관찰 자료들을 가져와 변수 tmp에 저장합니다
- 5줄 : 몸무게 변수만 선택하기
  - 변수 tmp는 변수 data와 마찬가지로 4개의 변수를 갖는 데이터 프레임으로, 여기서 세 번째 열을 가져와 변수 weight에 벡터로 저장합니다.
- 4줄과 5줄의 데이터 프레임에서 특정한 조건을 만족하는 자료 추출에 대해 이 장 마지막의 “다음 장을 위한 준비”에서 자세히 알아봅시다.

단일 모집단의 평균에 대한 가설검정

```
7.  barx <- mean(weight)
8.  s <- sd(weight)
9.  n <- length(weight)
10. h0 <- 2800
11. ( t.t <- (barx - h0) / (s / sqrt(n)) )
```

- **Step #2) 검정통계량 구하기**
  - 7~10줄 : 표본평균을 barx, 표본표준편차를 s, 표본의 개수를 n, 그리고 영가설 하에서의 평균(2800)을 h0에 저장합니다.
  - 11줄 : 위에서 구한 값들로 검정통계량을 계산하고, 변수 t.t에 저장한 후 출력합니다.

```
> ( t.t <- (barx - h0) / (s / sqrt(n)) )
[1] 2.233188
```

단일 모집단의 평균에 대한 가설검정

```
13. alpha <- 0.05
14. ( c.u <- qt(1-alpha, df=n-1) )
15. ( p.value <- 1 - pt(t.t, df=n-1) )
```

- **Step #3)** 판정을 위한 임계값과 유의확률 계산
  - 13줄 : 우리가 사용할 유의수준 0.05를 변수 alpha에 저장합니다.
  - 14줄 : 자유도가 n-1인 t-분포에서 $P(T > c_u) = 0.05$ 가 되는 임계값을 구해 변수 c.u에 저장하고 출력합니다.

```
> ( c.u <- qt(1-alpha, df=n-1) )
[1] 1.739607
```

  - 15줄 : 검정통계량이 저장되어 있는 변수 t.t를 이용하여 유의확률을 구해 변수 p.value에 저장하고, 값을 출력합니다.

```
> ( p.value <- 1 - pt(t.t, df=n-1) )
[1] 0.01963422
```

단일 모집단의 평균에 대한 가설검정

- 판정
  - 검정통계량(2.233)은 기각역(임계값 1.740)에 위치합니다.
  - 유의확률(0.0196)은 유의수준(0.05)보다 작습니다.
  - 위의 두 가지 방법 중 한가지를 선택하여 판정을 합니다.
    - 두 방법 모두 검정통계량을 이용하는 것으로 서로 같은 의미입니다.
    - 예의 자료에서는 영가설을 기각합니다.

단일 모집단의 평균에 대한 가설검정

```
17. t.test(weight, mu=2800, alternative="greater")
```

- **Step #4)** R에서의 단일표본의 평균 검정
  - R에서 t-분포를 이용한 검정은 t.test() 함수를 이용합니다.
  - 단일표본의 평균검정에서 사용할 경우 t.test() 함수는 다음과 같은 전달인자를 사용합니다.
    - 첫 번째 전달인자로 검정할 데이터를 전달합니다.
    - mu를 통해 전달되는 값은 영가설의 평균값입니다.
    - alternative 를 통해 전달되는 값은 대안가설에 맞춰 다음과 같이 지정합니다.

구분	대안가설	alternative
양쪽검정	$H_1 : \mu \neq \mu_0$	"two.sided" (기본값)
(왼쪽) 한쪽검정	$H_1 : \mu < \mu_0$	"less"
(오른쪽) 한쪽검정	$H_1 : \mu > \mu_0$	"greater"

단일 모집단의 평균에 대한 가설검정

- 17줄 : R의 t.test() 함수를 이용하여 앞에서 수립한 가설에 맞춰
  - 영가설에서 평균은 (mu=2800)이고
  - 대안가설은 평균이 2800보다 클 때(alternative="greater")입니다.

```
> t.test(weight, mu=2800, alternative="greater")

        One Sample t-test

data:  weight
① t = 2.2332, df = 17, p-value = 0.01963
② alternative hypothesis: true mean is greater than 2800
95 percent confidence interval:
 2873.477      Inf
③ sample estimates:
mean of x
 3132.444
```

단일 모집단의 평균에 대한 가설검정

- 단일 모집단의 평균 검정에 대한 t.test() 결과 해석하기

① 표본으로부터 구한 검정통계량, 자유도, 유의확률(p-value) 을 출력합니다.

- R은 사용자의 유의수준에 관심을 두지 않고 유의확률을 출력하여 사용자가 스스로 결정하도록 합니다. (SPSS 등 다른 통계 패키지에서도 이와 같이 출력합니다.)
- 검정통계량을 확인하고 이로부터 구한 유의확률을 통해 '판정'을 실시합니다.
  - 유의수준 0.05로 할 경우 유의확률 0.0196은 유의수준보다 작으므로 영가설을 기각합니다.

② 대안가설을 출력합니다.

- 앞서 alternative로 "greater"를 전달하였으므로 R은 대안가설로 평균이 2800보다 큰 경우로 하여 값을 계산하였음을 알려주고 있습니다.

③ 추정값을 출력합니다.

- 표본으로부터 구한 모평균에 대한 점추정 값과 95% 신뢰구간을 출력합니다.
- 신뢰수준을 90%로 변경하려면 t.test() 함수에 conf.level=0.90 과 같이 전달합니다.
- 우리의 예에서 신뢰구간에 영가설 하의 평균이 없습니다. 이와 같은 경우 대안가설을 선택하는 판정방법도 있습니다.

단일 모집단의 비율에 대한 가설검정

- **모비율에 대한 가설검정**
  - 5장에서 알아본 바와 같이 모비율은 우리가 관심을 갖고 있는 결과가 전체에서 얼마나 발생했는지를 비율로 나타낸 것으로,
  - 표본으로부터 표본비율을 구하고 이를 바탕으로 모비율에 대한 검정을 실시합니다.
  - 다음의 예제를 통해 모비율에 대한 검정 방법을 알아봅시다.
    - 다음의 예제는 제가 임의로 만든 자료로 실제 자료가 아님을 밝힙니다.

단일 모집단의 비율에 대한 가설검정

예제 3 단일 모집단의 비율 검정 : 야구공의 불량률 검정

KBO에서는 공인구를 납품받을 때 임의로 샘플을 추출하여 반발계수가 0.4134에서 0.4374 범위를 정상으로 인정하고, 정상 범위 바깥으로 발생하는 공이 10%를 넘기면 납품을 받지 않는다고 가정해봅니다. 공인구 제조사 A는 이 검사를 통과할 수 있을지 알아보기 위해, 사전에 납품하기 위해 준비한 공인구 중에서 100개를 표본으로 추출하여 반발계수를 조사한 결과를 6장 예제 파일의 data 폴더 아래에 restitution.txt로 우리에게 보내왔습니다. 이 자료로부터 공인구의 불량률이 10%를 넘는지 모비율에 대한 가설검정을 실시해봅시다.

- **5장 복습**
  - 모비율 P에 대한 추정량으로 표본비율 $\hat{p} = \frac{X}{n}$을 사용하고 표준오차 $\widehat{SE(\hat{p})} = \sqrt{\frac{n\hat{p}(1-\hat{p})}{n}}$임을 살펴보았습니다.

단일 모집단의 비율에 대한 가설검정

- **가설 수립**
  - 영가설 : 야구공 반발계수의 불량률은 10% 미만이다.
    - H_0 : $p_{불량} = 0.1$ 혹은 $p_{불량} < 0.1$
  - 대안가설 : 야구공 반발계수의 불량률은 10%를 넘는다.
    - H_0 : $p_{불량} \geq 0.1$
- **검정통계량 (영가설이 참이라는 가정하에 계산)**
  - 단일 모집단의 모비율 검정에서 사용하는 검정통계량

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0, 1^2)$$

- $\hat{p}$: 표본비율, p_0 : 영가설 하의 모비율, n : 표본의 개수, $\sqrt{\frac{p_0(1-p_0)}{n}}$: 영가설 하의 표준오차

단일 모집단의 비율에 대한 가설검정

- 검정통계량 계산
  - 검정통계량을 구하고자 표본비율을 계산하기 위해 야구공의 반발계수가 정상 영역이 아닌, 0.4134보다 작거나 0.4374보다 크게 관찰된 공의 수를 구합니다.
    - 예제파일에서는 전체 100중 11개가 불량으로 조사되어 $\hat{p} = \frac{11}{100} = 0.11$ 입니다.
  - 영가설 하에서의 불량률은 0.1이고, 표본의 수는 100개로 검정통계량은 다음과 같습니다.

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.11 - 0.1}{\sqrt{\frac{0.1 \times 0.9}{100}}} \cong 0.333$$

단일 모집단의 비율에 대한 가설검정

- 유의수준 : $\alpha = 0.05$
  - 검정통계량이 따르는 분포는 표준정규분포입니다.
  - 대안가설에 의해 '(오른쪽) 한쪽검정'으로 임계값은 $P(T > c_u) = 0.05$가 되는 c_u로, R을 통해 약 1.645 임을 계산할 수 있습니다.

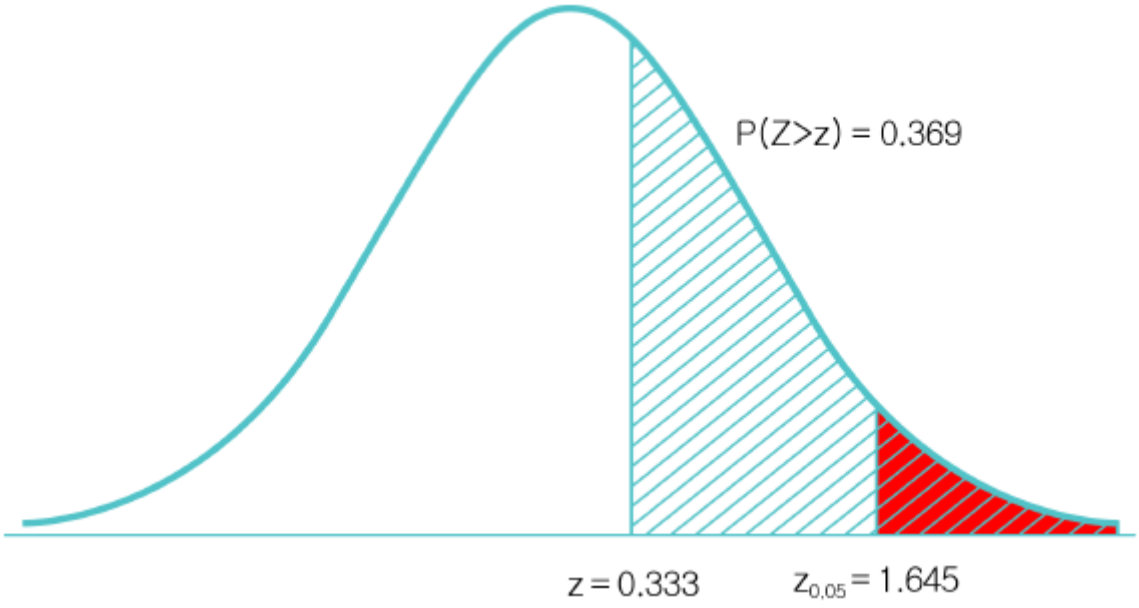
```
> alpha <- 0.05
> ( c.u <- qnorm(1-alpha) )
[1] 1.644854
```

  - 검정통계량에 대한 유의확률($P(Z > 0.333)$)은 R을 통해 약 0.369임을 알 수 있습니다

```
> ( p.value <- 1 - pnorm(z) )
[1] 0.3694413
```

단일 모집단의 비율에 대한 가설검정

- **판정**
  - 기각역을 이용한 판정
    - (오른쪽) 임계값은 1.645로 검정통계량 0.333이 채택역에 위치합니다. 이로부터 영가설을 채택합니다.
  - 유의확률(p-value)을 이용한 판정
    - 유의확률은 0.369로 유의수준 0.05보다 큰 값입니다. 이로부터 영가설을 채택합니다.



단일 모집단의 비율에 대한 가설검정

- **결론**
  - 생산된 야구공의 불량률이 10%를 넘는지를 알아보기 위해 제품 중 무작위로 100개의 공을 추출하여 반발계수를 유의수준 0.05에서 가설검정을 실시한 결과,
  - 11개의 공에서 반발계수가 0.4134보다 작거나 0.4374보다 크게 관찰되었으며,
  - 불량률에 대한 검정통계량은 0.333(유의확률 0.369)으로 나타났습니다.
  - 이로부터 "야구공의 불량률은 10% 미만"이라는 영가설을 기각할 수 없었습니다.
  - 공장에서 생산된 야구공의 불량률은 10% 미만으로 잘 유지되고 있다고 할 수 있습니다.
  - 생각해 볼 문제
    - 무작위 추출에서 공의 불량이 100개 중 10개를 넘어선 경우로 가설검정에서는 이와 같은 결론을 얻을 수 있습니다.
    - 하나의 생산단위에서 무작위로 100개를 추출하여 10개의 불량을 넘을 경우 해당 생산단위에서 생산된 제품 전체를 폐기하기도 합니다.

단일 모집단의 비율에 대한 가설검정

예제 6-2 모비율 검정 : 야구공의 불량률 검정

준비파일 | 06.pop.proportion.R

- **R을 이용하여 모비율을 검정합니다.**
  - 먼저 R 코드를 이용하여 검정통계량과 유의확률을 직접 구합니다.
  - R에서 제공하는 함수를 이용하여 검정합니다.
    - 표준정규분포를 이용한 검정이 아닌 다른 분포를 이용하지만, 동일한 결과를 얻음을 확인합니다.

단일 모집단의 비율에 대한 가설검정

```
1. tmp <- read.table("./data/restitution.txt", header=T)
2. rel <- ifelse(tmp$rst < 0.4134 | tmp$rst > 0.4374, 1, 0)
```

- **Step #1) 데이터 준비하기**
  - 1줄 : Chapter06 하위의 data 디렉토리에서 restitution.txt를 읽어 변수 tmp에 데이터 프레임으로 저장합니다.
    - 해당 파일에서 첫 줄은 데이터 프레임의 열의 이름(변수명)으로 인식하도록 합니다 (header=T).
  - 2줄 : 읽어온 데이터는 rst 열 하나만을 포함하고 있으며, 표본으로 추출된 각 공의 반발계수를 저장하고 있습니다. 이 데이터로 부터
    - 반발계수가 0.4134보다 작거나 0.4374보다 크면 1, 그렇지 않으면 0으로 구성된 벡터를 생성하고 변수 rel에 저장합니다. (불량이면 1, 정상이면 0이 되도록 합니다)
    - 이 장 뒷부분의 7장을 위한 준비를 참고해 주세요.

단일 모집단의 비율에 대한 가설검정

```
4.  n <- length(rel)
5.  nos <- sum(rel)
6.  sp <- nos / n
7.  hp <- 0.1
8.  (z <- (sp - hp) / sqrt( ( hp*(1-hp) )/n ) )
```

- **Step #2) 검정통계량 구하기**
  - 4~7줄 : 검정통계량을 구하기 위해 필요한 표본의 개수(n), 불량품의 개수(nos), 표본의 불량률(sp), 영가설 하의 모비율(hp)을 구합니다.
  - 5줄 : 위에서 생성한 벡터 rel에서 1은 불량품, 0은 정상 제품을 나타내므로, rel의 합을 통해 불량품의 개수를 구할 수 있습니다.
  - 8줄 : 검정통계량을 계산하고 출력합니다

```
> (z <- (sp - hp) / sqrt( ( hp*(1-hp) )/n ) )
[1] 0.3333333
```

단일 모집단의 비율에 대한 가설검정

```
10. alpha <- 0.05
11. ( c.u <- qnorm(1-alpha) )
12. ( p.value <- 1 - pnorm(z) )
```

- **Step #3) 판정을 위한 임계값과 유의확률 계산 및 출력**
  - 10줄 : 유의수준 0.05를 변수 alpha에 저장합니다.
  - 11줄 : 표준정규분포에서 $P(T > c_u) = 0.05$ 가 되는 임계값을 구해 변수 c.u 에 저장하고 출력합니다.

```
> ( c.u <- qnorm(1-alpha) )
[1] 1.644854
```

  - 12줄 : 검정통계량이 저장되어 있는 변수 z를 이용하여 유의확률을 구해 변수 p.value에 저장하고 값을 출력합니다. 유의확률이 유의수준보다 큽니다

```
> ( p.value <- 1 - pnorm(z) )
[1] 0.3746353
```

단일 모집단의 비율에 대한 가설검정

- 판정
  - 검정통계량(0.333)은 채택역(임계값 1.645)에 위치합니다.
  - 유의확률(0.3746353)은 유의수준(0.05)보다 큽니다.
  - 위의 두 가지 방법 중 한가지를 선택하여 판정을 합니다.
    - 두 방법 모두 검정통계량을 이용하는 것으로 서로 같은 의미입니다.
    - 예의 자료에서는 영가설을 채택합니다.

단일 모집단의 비율에 대한 가설검정

```
14. prop.test( nos, n, p=0.1,
                  alternative="greater", correct=FALSE)
```

- **Step #4) R에서의 단일표본의 모비율 검정**
  - R에서 단일표본의 모비율 검정은 prop.test() 함수를 이용합니다.
    - 첫 번째 전달인자로 성공의 횟수를 전달합니다.
    - 두 번째 전달인자로 전체 표본의 개수를 전달합니다.
    - p를 통해 영가설 하의 모비율을 전달합니다.
    - alternative를 통해 모평균의 가설검정에서와 같이 검정의 방법을 전달합니다.
    - correct를 통해 이산형 자료를 연속형 분포로 검정하는 데에 따른 보정을 위한 Yates 연속성 수정의 여부로 우리가 사용한 검정통계량을 사용하기 위해 본 예에서는 FALSE를 전달합니다.

단일 모집단의 비율에 대한 가설검정

```
> prop.test(nos, n, p=0.1, alternative="greater", correct=FALSE)

        1-sample proportions test without continuity correction

① data:  nos out of n, null probability 0.1
② X-squared = 0.11111, df = 1, p-value = 0.3694
③ alternative hypothesis: true p is greater than 0.1
④ 95 percent confidence interval:
 0.0684615 1.0000000
sample estimates:
   p
0.11
```

단일 모집단의 비율에 대한 가설검정

- 단일 모집단의 모비율 검정에 대한 prop.test() 결과 해석하기

① 영가설을 출력합니다.

② 표본으로부터 구한 검정통계량, 자유도, 유의확률(p-value) 을 출력합니다.

- 검정통계량이 표준정규분포를 따르는 Z가 아닌 자유도가 1인 χ^2-분포를 따르는 통계량임을 알 수 있습니다.
- 자유도가 1인 χ^2-분포는 앞서 학습한 바와 같이 하나의 표준정규분포를 제곱한 것으로 prop.test()가 구한 검정통계량 0.11111의 제곱근, 즉 $\sqrt{0.11111} = 0.33333$으로 앞서 구한 검정통계량 Z와 같습니다.
- 유의수준 0.05로 할 경우 유의확률 0.3694은 유의수준보다 크므로 영가설을 채택합니다.

③ 대안가설을 출력합니다.

- 앞서 alternative로 “greater”를 전달하였으므로 R은 대안가설로 모비율 0.1보다 큰 경우로 하여 값을 계산하였음을 알려주고 있습니다.

④ 추정값을 출력합니다.

- 표본으로부터 구한 모평균에 대한 점추정 값과 95% 신뢰구간을 출력합니다.

7장을 위한 준비

: 데이터 프레임 다루기와 데이터 정제연습

데이터 프레임 다루기

- **데이터 프레임 : 데이터 셋**
  - 거의 모든 통계자료는 데이터 프레임의 형태를 갖고 있으며, 이는 우리가 일반적으로 보는 표 형태의 자료 집합(데이터 셋)입니다.
    - 자료들의 모임은 행과 열로 되어 있고, 각 열에 관찰대상(행)들로부터 관찰한 속성(변수)들이 위치합니다.
    - 각 속성은 벡터 혹은 factor 형태의 자료이며 모든 속성(변수)들의 크기는 동일해야 합니다.
  - 앞서 살펴본 신생아 자료는 4개의 변수를 44개의 관찰대상으로부터 관찰한 데이터 프레임입니다.
  - 이 자료를 이용해 데이터 프레임을 다루는 방법을 조금 더 알아보겠습니다.

```
data <- read.table
("http://www.amstat.org/publications/jse/datasets/babyboom.dat.txt",
header=F)
```

데이터 프레임의 정보 취득하기

- **nrow()와 ncol()을 이용한 행의 수와 열의 수 확인**
  - nrow() : 데이터 프레임을 구성하는 자료의 행의 수 출력(관찰대상 수, 표본수)
  - ncol() : 데이터 프레임을 구성하는 자료의 열의 수 출력(변수의 개수)

```
> nrow(data)
[1] 44
> ncol(data)
[1] 4
```

- **str()을 이용한 구조 확인 (structure)**
  - R은 다양한 자료들의 모임(벡터, 데이터프레임, 행렬, 리스트, factor 등)을 지원합니다.
  - str() 함수는 이런 자료들의 모임이 어떤 구조로 되어 있는지 확인하는 함수로 이를 이용하여 읽어온 자료의 구조를 확인합니다.
    - 앞서 사용했지만 데이터 프레임의 경우에 대해 조금 더 살펴보겠습니다.

데이터 프레임 다루기

자료들의 모임인 자료구조의 유형 출력
: 데이터 프레임은 “data.frame”

데이터 프레임의 경우 관찰대상수와 변수의 수 출력

```
> str( data )
'data.frame':   44 obs. of  4 variables:
 $ V1: int  5 104 118 155...
 $ V2: int  1 1 2 2 ...
 $ V3: int  3837 3334 3554 3838 ...
 $ V4: int  5 64 78 115 ...
```

데이터 프레임을 구성하는 변수 출력
- $ 표시 이후 변수명 출력

각 구성변수의 자료형과 미리보기
- 벡터의 경우 위와 같이 벡터를 이루는 자료형을 출력합니다.
- factor의 경우 factor를 이루는 수준을 출력합니다.

데이터 프레임 다루기

- **head()와 tail()을 통해 자료의 앞부분과 뒷부분 확인**
  - 자료를 불러오면 제대로 불러왔는지
    - str() 함수를 통해 차원(행과 열의 수)을 확인하고,
    - 각 열의 자료형을 탐색하며,
    - 추가적으로 자료의 앞부분과 뒷부분을 읽어 잘 불러왔는지 확인합니다.
      - head() 는 자료의 앞부분을 tail() 은 자료의 뒷부분 일부를 불러옵니다.

```
> head(data)
   V1 V2   V3  V4
1   5  1 3837   5
2 104  1 3334  64
3 118  2 3554  78
4 155  2 3838 115
5 257  2 3625 177
6 405  1 2208 245
```

```
> tail(data)
   V1 V2   V3   V4
39 2051  2 3370 1251
40 2104  2 2121 1264
41 2123  2 3150 1283
42 2217  1 3866 1337
43 2327  1 3542 1407
44 2355  1 3278 1435
```

데이터 프레임 다루기

- head()와 tail()은 첫 번째 전달인자로 처음 혹은 끝을 확인할 자료의 이름을 전달받아 각각 여섯 개의 자료를 출력합니다.
  - 다음과 같이 n을 통해 확인하고자 하는 자료의 개수를 지정하여 살펴 볼 수 있습니다.

```
> head(data, n=2)
   V1 V2   V3 V4
1   5  1 3837  5
2 104  1 3334 64

> tail(data, n=3)
     V1 V2   V3   V4
42 2217  1 3866 1337
43 2327  1 3542 1407
44 2355  1 3278 1435
```

- 자료를 읽어올 때 모든 자료를 살펴보는 것이 좋지만, 그 수가 많은 경우에는 이렇게 일부를 (특히 뒷부분) 확인하여 잘 불러왔는지 확인합니다.

데이터 프레임 다루기

- **names()를 이용한 열의 이름 확인 및 변경**
  - names() 함수를 이용하면 데이터 프레임의 각 열의 이름을 확인할 수 있습니다.
  - names() 함수에 전달인자로 알고자 하는 데이터 프레임을 저장하고 있는 변수명을 입력하면 다음과 같이 출력됩니다.

```
> names(data)
[1] "V1" "V2" "V3" "V4"
```

  - 데이터 프레임의 열 이름을 변경할 수 있습니다.
    - names()가 문자열 벡터를 저장하고 있는 것을 사용자가 원하는 벡터로 변경하는 과정입니다.

데이터 프레임 다루기

```
1. > names(data) <- c("time", "gender", "weight", "minutes")
2. > names(data)
   [1] "time"    "gender"  "weight"  "minutes"
3. > names(data)[1] <- "time.24Hrs"
4. > names(data)
   [1] "time.24Hrs" "gender"     "weight"     "minutes"
```

- 1줄 : names(data)를 문자열 벡터 c("time", "gender", "weight", "minutes")로 바꿉니다.
- 2줄 : 변경된 내용을 확인해봅니다. 모든 변수의 이름이 바뀌었습니다.
- 3줄 : names(data)의 결과는 벡터로 names(data)[1]은 names(data) 벡터의 첫 번째 원소를 나타내며, 그 값을 "time.24Hrs"로 변경합니다.
- 4줄 : 변경된 내용을 확인해봅니다. 첫 번째 원소의 값만 "time.24Hrs"로 변경되었습니다.

데이터 프레임 다루기

- 데이터 프레임에서 row.names()는 행 번호(이름)를 나타내는 함수로 데이터에 그 값이 포함되어 있지 않더라도 R의 데이터 프레임은 자료의 정보로 알아서 생성합니다.
  - 만일 행의 번호가 아닌 이름을 주고 싶은 경우 변경할 수 있습니다.

```
> row.names(data)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
[11] "11" "12" "13" "14" "15" "16" "17" "18" "19" "20"
[21] "21" "22" "23" "24" "25" "26" "27" "28" "29" "30"
[31] "31" "32" "33" "34" "35" "36" "37" "38" "39" "40"
[41] "41" "42" "43" "44"
```

  - 만일 변경하고자 하는 경우 다음과 같이 변경할 수 있습니다.
    - 각 행의 이름으로 “Row #행번호”로 바꾸는 경우

```
> row.names(data) <- paste("Row #", row.names(data), sep="")
```

데이터 프레임 다루기

- **추출하고자 하는 열 선택하기**
  - 다음은 다양한 방법으로 데이터 프레임 data에서 gender 열을 추출하는 방법입니다.
    - 각 방법의 결과로 가져오는 열이 벡터인지 데이터 프레임인지 잘 확인해 주시기 바랍니다.
    - str() 함수로 벡터의 구조를 확인하는 경우에는 “자료형 [1:전체개수] 미리보기” 형태로 출력됩니다.

데이터 프레임 다루기

```
> g1 <- data$gender                                    ①
> str(g1)
 int [1:44] 1 1 2 2 2 1 1 2 2 2 ...
> g2 <- data[,2]                                       ②
> str(g2)
 int [1:44] 1 1 2 2 2 1 1 2 2 2 ...
> g3 <- data["gender"]                                 ③
> str(g3)
'data.frame':  44 obs. of  1 variable:
 $ gender: int  1 1 2 2 2 1 1 2 2 2 ...
> g4 <- data[[2]]                                      ④
> str(g4)
 int [1:44] 1 1 2 2 2 1 1 2 2 2 ...
> g5 <- data[["gender"]]
> str(g5)
 int [1:44] 1 1 2 2 2 1 1 2 2 2 ...
```

데이터 프레임 다루기

- $를 이용한 열 지정
  - 데이터 프레임 이름 뒤에 $ 표시 후에 열 이름을 넣으면 해당 열의 모든 자료를 벡터로 추출해줍니다.
- [ ]를 이용한 열 지정
  - [ ]를 이용할 때 콤마 앞은 행번호를, 콤마 뒤는 열번호를 지정합니다.
  - 만일 그 값이 비워져 있으면 모든 행 또는모든 열을 의미합니다.
  - 예에서 [, 2]로입력하여 두 번째 열의 모든 행의 자료를 추출합니다.
  - 추출된 형태는 벡터입니다.
- ["변수명"]을 이용한 열 지정
  - [ ] 안에 문자열로 변수명을 입력하면 데이터 프레임으로 해당 자료를 추출합니다.
- [[ ]]을 이용한 열 지정
  - 두 개의 대괄호 [[ ]]를 사용하는 경우 열 번호와 열 이름의 문자열 모두를 전달할 수 있으며 해당열의모든 행의 자료를 추출해줍니다. 결과물은 벡터입니다.

데이터 프레임 다루기

- 다음과 같이 2개 이상의 열을 가져올 수 있으며, 이 경우에는 모두 데이터 프레임의 형태로 추출해줍니다.
  - 가져오고자 하는 열의 순서 혹은 이름을 벡터로 지정합니다.

```
> gg1 <- data[, c(2, 4)]
> str( gg1 )
'data.frame':  44 obs. of  2 variables:
 $ gender : int  1 1 2 2 2 1 1 2 2 2 ...
 $ minutes: int  5 64 78 115 177 245 247 262 271 428 ...

> gg2 <- data[c("gender", "minutes")]
> str( gg2 )
'data.frame':  44 obs. of  2 variables:
 $ gender : int  1 1 2 2 2 1 1 2 2 2 ...
 $ minutes: int  5 64 78 115 177 245 247 262 271 428 ...
```

데이터 프레임 다루기

- **조건에 맞는 행 선택하기**
  - 조건에 맞는 행 선택하기는 대괄호([ ])를 이용한 직접 지정방법과 앞서 사용한 subset() 함수를 이용하는 두 가지 방법이 있습니다.
  - 예제 1) 남아 신생아의 자료 가져오기 (남아 신생아는 gender 값이 2입니다.)
    - 직접 지정 : data[data$gender==2,]
    - 함수 이용 : subset(data, gender==2)

```
> str( data[data$gender==2, ] )
'data.frame':  26 obs. of  4 variables:
 $ time.24Hrs: int  118 155 257 422 431 ...
                          ...
 $ minutes   : int  78 115 177 262 271 ...

> str( subset(data, gender==2) )
'data.frame':  26 obs. of  4 variables:
 $ time.24Hrs: int  118 155 257 422 431 ...
                          ...
 $ minutes   : int  78 115 177 262 271 ...
```

데이터 프레임 다루기

- 예제 2) 남아 신생아 자료에서 평균 체중보다 큰 자료만 가져오기
  - 행을 선택하는 조건이 두 가지로 늘어났습니다. (남아이고 평균체중보다 큰 신생아)
  - 각 조건을 '&' 연산자로 결합하여 두 조건을 만족하는 행을 출력합니다.
    - 직접 지정 : data[data$gender==2 & data$weight > male.m,]
    - 함수 이용 : subset(data, (gender==2) & (weight > male.m))

```
> male.m <- mean(data$weight)
> str( data[data$gender==2 & data$weight > male.m, ] )
'data.frame':  19 obs. of  4 variables:
 $ time.24Hrs: int  118 155 257 708 735 ...
                                  ...
 $ minutes   : int  78 115 177 428 455 ...

> str( subset(data, (gender==2) & (weight > male.m) ) )
'data.frame':  19 obs. of  4 variables:
 $ time.24Hrs: int  118 155 257 708 735 ...
                                  ...
 $ minutes   : int  78 115 177 428 455 ...
```

데이터 프레임 다루기

- **조건에 맞는 행과 열 선택하기**
  - 열 선택과 행 선택을 이용해
    - 남아 신생아들의 자료 중에 평균 체중보다 큰 아이들의 체중과 (행 선택 조건)
    - 출생시간의 분(24시를 기준으로 분으로 측정, 열 선택 조건)만 가져와봅시다.
    - 대괄호를 이용한 직접 지정은 원하는 행, 원하는 열을 대괄호 안에 직접 지정하여 가져오고,
      - data[data$gender==2 & data$weight > male.m, **c(2, 4)**]
    - subset() 함수는 조건에 맞는 행들로 구성된 부분집합을 만들고 특정 열을 선택하기 위해 별도의 전달인자(select)를 이용하여 가져옵니다.
      - subset(data, (gender==2) & (weight > male.m), **select=c(2, 4)**)

데이터 프레임 다루기

```
> str( data[data$gender==2 & data$weight > male.m, c(2, 4)] )
'data.frame':  19 obs. of  2 variables:
 $ gender : int  2 2 2 2 2 2 2 2 2 2 ...
 $ minutes: int  78 115 177 428 455 492 635 776 785 914 ...

> str(subset(data, (gender==2) & (weight > male.m), select=c(2, 4)))
'data.frame':  19 obs. of  2 variables:
 $ gender : int  2 2 2 2 2 2 2 2 2 2 ...
 $ minutes: int  78 115 177 428 455 492 635 776 785 914 ...
```

데이터 프레임 다루기

- **데이터 프레임 저장하기**
  - write.table() 함수를 이용해 자료를 파일로 저장해 봅시다.
  - 자주 쓰이는 write.table() 함수가 필요로 하는 전달인자는 다음과 같습니다.

전달인자	설명	예시
첫 번째(x)	저장할 데이터 프레임의 이름	data 혹은 x=data
두 번째(file)	저장할 파일의 경로와 이름	"./data/sample3.txt"
row.names	행 이름(행 번호)의 저장 여부 기본값은 TRUE	row.names=FALSE 행 번호는 저장에서 제외할 때
col.names	열 이름의 저장 여부, 기본값은 TRUE	col.names=FALSE 열 이름을 저장에서 제외할 때
sep	열 구분자, 기본값은 공백문자 " "	sep="," csv와 같이 콤마를 열 구분자로 사용할 경우, 미지정 시 공백문자 " "
na	결측값으로 저장할 문자열, 기본값은 "NA"	na="9999" 결측값을 9999로 저장하려면, 미지정 시 "NA"
append	기존 파일의 뒤에 붙일 것인지 여부, 기본값은 FALSE이며 이 경우 기존 파일 위에 쓴다.	append=TRUE 기존 파일 뒤에 결과를 붙일 경우 주로 로그 파일 등에서 사용하며, 기본값은 FALSE

데이터 프레임 다루기

- 다음과 같이 data 데이터 프레임의 2열과 3열로 구성된 데이터 프레임 변수 chapter7을 data 디렉토리 아래에 chapter7.txt(./data/chapter7.txt)로 저장해 봅시다.

```
> chapter7 <- data[, c(2, 3)]
> write.table(chapter7, "./data/chapter7.txt")
```

  - 저장한 파일을 메모장 같은 포로그램으로 확인해보면 다음과 같이 행 번호가 문자열로 저장된 것을 확인할 수 있습니다. (열 구분자는 한 개의 빈칸 " ")

```
1 "gender" "weight"
2 "1"  1  3837
3 "2"  1  3334
4 "3"  2  3554
5 "4"  2  3838
6 "5"  2  3625
7 "6"  1  2208
8 "7"  1  1745
```

데이터 프레임 다루기

- 다음과 같이 row.names=FALSE를 통해 행 번호를 제거한 결과를 저장합니다.
- 이 파일을 7장에서 사용합니다.
  - 7장을 위한 프로젝트를 만들고 data 폴더를 만들어 복사해 주세요.

```
> write.table(chapter7, "./data/chapter7.txt", row.names=FALSE)
```

  - append를 통해 값을 전달하지 않아 기본값인 FALSE가 되어 동일한 파일이름을 갖는 파일이 있을 때 기존 파일을 덮어 쓰게 됩니다.
  - sep 또한 마찬가지로 지정하지 않아 기본값인 공백문자 한 개(" ")를 이용하여 각 열을 구분합니다.

```
1 "gender" "weight"
2 1 3837
3 1 3334
4 2 3554
5 2 3838
6 2 3625
7 1 2208
8 1 1745
```

데이터 정제하기 연습

- **7장에서 사용할 데이터 파일**
  - 프로젝트 디렉토리 'Chapter06/data/'에 있는 데이터 파일 'age.data.csv'
    - 실제 자료가 아닌, 가상의 자료입니다.
  - 거주지역의 규모별로 50명씩 표본을 추출하여 성별, 서비스 평가점수, 나이를 측정한 자료로 구조는 다음과 같습니다.

변수명	저장 자료형	변수 설명	자료 설명
scale	숫자형	거주지역의 규모 (범주형)	1 : 특별시, 광역시 2 : 시 지역 3 : 읍면 지역
sex	숫자형	성별(범주형)	1 : 여성 2 : 남성
score	숫자형	서비스 평가점수	0부터 10까지의 서비스 평가점수로 높을수록 좋은 서비스
age	숫자형	나이	

  - 다음과 같이 파일을 읽어오고 원하는 구조를 유지하는지 확인해 봅시다.

데이터 정제하기 연습

```
1.  ad <- read.csv("./data/age.data.csv", header=T)
2.  str( ad )
3.  head( ad )
4.  tail( ad )
5.  summary(ad)
```

- **Step #1) 파일을 읽고 원하는 구조를 갖추고 있는지 확인합니다.**
  - 1줄 : 불러올 파일은 './data/age.data.csv'에 있으며, 첫 줄은 변수명으로 되어 있으므로 데이터로 읽지 않고 변수명으로 읽고 이를 ad에 저장합니다.
  - 2줄 : str() 함수를 이용해서 전체 구조를 살펴봅니다.
    - 150개의 관찰 자료로부터 4개의 변수를 읽어왔으며, 모든 변수는 숫자형입니다.

```
'data.frame':  150 obs. of  4 variables:
 $ scale: int  1 1 1 1 1 1 1 1 1 1 ...
 $ sex  : int  2 2 2 1 1 2 1 2 2 2 ...
 $ score: int  8 5 7 4 5 3 3 7 9 4 ...
 $ age  : int  56 33 49 53 74 42 51 59 25 57 ...
```

데이터 정제하기 연습

- 3, 4줄 : 관찰 자료가 많지 않다면 한눈에 살필 수 있으나, 많을 경우에는 자료의 앞과 뒤를 읽어 자료의 구조(변수 등)가 잘못되지 않고 자료를 잘 읽었는지 확인합니다.
  - 현재 읽어온 데이터는 마지막에 변수로써 판별이 잘못되어 값이 밀려서 나오는 등 구조상으로는 큰 문제가 없어 보입니다.

```
> head( ad )
  scale sex score age
1     1   2     8  56
2     1   2     5  33
3     1   2     7  49
4     1   1     4  53
5     1   1     5  74
6     1   2     3  42
```

```
> tail(data)
    scale sex score age
145     3   1     6  62
146     3   2     6  33
147     3   2     7  54
148     3   2     6  61
149     3   1     8  46
150     3   1     4  15
```

데이터 정제하기 연습

- 5줄 : summary() 함수를 사용하여 각 변수의 요약통계를 확인합니다.
  - 변수가 숫자 자료로 구성될 경우 사분위수와 평균을 보여주어 대략적인 모습을 확인할 수 있도록 합니다.
  - 또한 범주형 자료일 경우에는 각 범주에 해당하는 개수를 세어줍니다
  - 결측 자료의 개수를 세어줍니다.

```
> summary(ad)
     scale           sex             score             age        
 Min.   :1   Min.   :1.000   Min.   : 1.00   Min.   :14.00  
 1st Qu.:1   1st Qu.:1.000   1st Qu.: 4.00   1st Qu.:38.00  
 Median :2   Median :2.000   Median : 6.00   Median :46.00  
 Mean   :2   Mean   :1.507   Mean   : 8.22   Mean   :46.51  
 3rd Qu.:3   3rd Qu.:2.000   3rd Qu.: 7.00   3rd Qu.:56.00  
 Max.   :3   Max.   :2.000   Max.   :99.00   Max.   :89.00  
```

데이터 정제하기 연습

- 각 변수들의 요약통계를 보면 0부터 10까지 측정한 서비스 점수의 최댓값이 99로 조사되었음을 확인할 수 있습니다.
  - 여기서 99는 결측값을 나타내는 값으로 전통적인 방식입니다.
  - 이 방식은 설문지에 기록된 서비스 점수를 파일로 작성하는 과정(코딩)에서 응답을 하지 않았을 경우나 응답이 불분명하여 결측으로 판단한 경우등을 나타내기 위해 사용합니다.
    - 서비스 점수의 경우 관측할 수 있는 값이 두 자릿수인 10으로, 두 자리 숫자 중 가장 큰 값인 99를 결측을 나타내는 값으로 사용합니다
  - 만일 결측치를 나타내는 99와 같은 값을 NA로 변경하지 않으면 잘못된 통계를 도출할 수 있으므로 자료를 가져온 후 올바른 분석을 위해 반드시 처리해야 합니다.
  - 이에 결측으로 기록된 99를 R에서 결측으로 사용하는 NA로 바꿔보겠습니다.

데이터 정제하기 연습

```
7. ad$score <- ifelse(ad$score==99, NA, ad$score)
8. summary(ad)
```

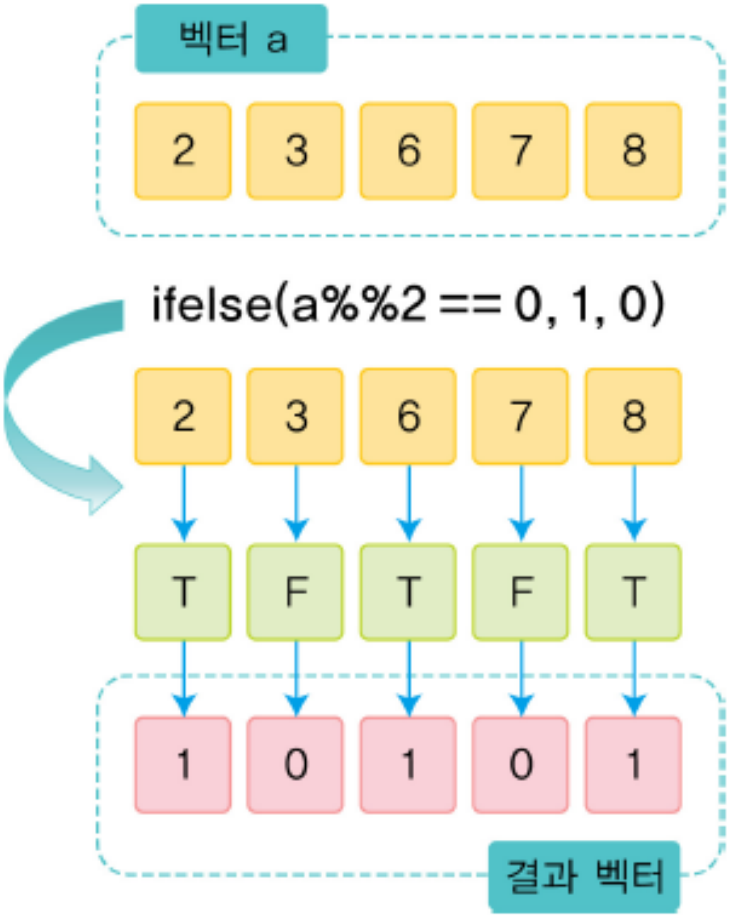
- 7줄 : ad$score의 각 값을 99와 같은지 비교하여 99이면 NA를, 그렇지 않으면 해당 원소의 기존 값(ad$score)을 갖는 결과 벡터를 만든 후, 이 벡터로 ad$score를 변경합니다.
- 8줄 : summary() 함수를 이용해서 확인합니다.
  - 최댓값이 10이 되었고 결측값은 4개가 있음을 알 수 있습니다

```
scale          sex             score             age
Min.   :1   Min.   :1.000   Min.   : 1.000   Min.   :14.00
1st Qu.:1   1st Qu.:1.000   1st Qu.: 4.000   1st Qu.:38.00
Median :2   Median :2.000   Median : 6.000   Median :46.00
Mean   :2   Mean   :1.507   Mean   : 5.733   Mean   :46.51
3rd Qu.:3   3rd Qu.:2.000   3rd Qu.: 7.000   3rd Qu.:56.00
Max.   :3   Max.   :2.000   Max.   :10.000   Max.   :89.00
                            NA's   :4
```

데이터 정제하기 연습

- **ifelse() 함수**
  - 조건을 처리하는 함수로, 다음과 같이 구성되어 결과로 벡터를 반환합니다.
    - 첫 번째 전달인자로 비교하고자 하는 벡터의 논리연산(TRUE와 FALSE로 구성된 벡터를 반환하는 연산)을 전달합니다.
    - 두 번째 전달인자로 첫 번째 전달인자로 전달된 논리연산이 참일 때 반환할 값
    - 세 번째 전달인자로 첫 번째 전달인자로 전달된 논리연산이 거짓일 때 반환할 값
  - 반환되는 벡터는
    - 첫 번째 전달인자를 통해 비교하고자 하는 벡터의 각 원소별로
    - 논리연산이 TRUE일 경우 두 번째 전달인자로 전달된 값을,
    - 논리연산이 FALSE일 경우 세 번째 전달인자로 사용된 값을 갖는 벡터입니다.
  - 예제) (2, 3, 6, 7, 8)로 구성된 백터의 홀수/짝수 구분
    - 짝수인 값을 1, 그렇지 않은 값(홀수인 값)을 0으로 하는 벡터를 ifelse()를 이용하여 생성하는 과정은 다음 그림과 같습니다.

데이터 정제하기 연습



a의 각 원소별로

2로 나눈 값이 0이면 TRUE

TRUE인 경우 1, 그렇지 않으면 0으로 구성된 벡터 반환

데이터 정제하기 연습

- **결측값을 고려한 자료 불러오기**
  - 앞선 자료에서 결측값은 99로 코딩된 사실을 미리 알고 있었다면, 파일을 불러오는 함수에서 이를 처리할 수 있습니다.

```
10. ad2 <- read.csv("./data/age.data.csv", header=T,
                    na.strings=c("99"))
11. summary(ad2)
```

  - 11줄 : read.csv() 혹은 read.table() 등 외부 파일을 읽어오는 함수에서 사용하는 na.strings는 결측값으로 사용할 문자열을 지정하여 해당 문자들을 NA로 읽게 합니다.
    - 각 열의 데이터 중 문자열 "99" 를 결측값으로 인식하도록 하였습니다. (파일에서 읽는 것으로 문자열 "99"를 전달하더라도 R이 결측값 외에 다른 자료들이 숫자로만 구성된 자료라면 알아서 숫자형으로 읽어옵니다.)
    - na.strings는 결측값으로 처리할 문자들을 벡터로 받아서 처리하는데 이는 여러 개의 문자들을 결측값으로 처리할 수 있음을 뜻합니다.(예. c("99", "-", "!NULL"))
  - 12줄 : 앞서 결측값을 따로 처리한 결과와 비교해 봅시다.

데이터 정제하기 연습

```
> summary(ad2)
     scale          sex            score            age       
 Min.   :1   Min.   :1.000   Min.   : 1.000   Min.   :14.00  
 1st Qu.:1   1st Qu.:1.000   1st Qu.: 4.000   1st Qu.:38.00  
 Median :2   Median :2.000   Median : 6.000   Median :46.00  
 Mean   :2   Mean   :1.507   Mean   : 5.733   Mean   :46.51  
 3rd Qu.:3   3rd Qu.:2.000   3rd Qu.: 7.000   3rd Qu.:56.00  
 Max.   :3   Max.   :2.000   Max.   :10.000   Max.   :89.00  
                             NA's   :4
```

데이터 정제하기 연습

- 결측값이 있는 서비스 만족도 점수의 평균을 구해 봅시다.
  - 2장에서 na.rm=TRUE를 이용하여 결측값이 있는 자료의 평균을 구해 봤습니다.
  - 결측값을 처리하는 방법을 익히기 위해 조금 더 알아봅시다.

```
> mean(ad$score)
[1] NA
> mean(ad$score, na.rm=TRUE)
[1] 5.732877
```

  - 하나라도 결측값이 있는 자료에 대해 mean()과 같은 함수를 적용하면 NA로 나옵니다. 이는 벡터내의숫자를 이용하여 집계를 내는 함수들에 대해서는 동일하게 작용합니다(var(), sd(), median() 등)
    - length()의 경우 계산을 하는 함수가 아니므로 영향을 받지 않습니다.
  - 결측값이 있는 자료의 평균을 구할 때 na.rm을 이용하여 TRUE를 전달하면, 결측값을 제외한 나머지 값들의 평균을 구합니다
  - R에서 결측값을 다루는 방법에 대해 알아봅시다.

데이터 정제하기 연습

- **결측 판별 함수 is.na()**
  - 결측 자료 처리를 위해 결측값을 판별하는 함수 is.na()를 사용합니다.
    - is.na()는 주어진 자료가 NA이면 TRUE를, 그렇지 않으면 FALSE를 반환하는 함수입니다.

```
16. is.na( c(1, NA, 3, NA, 5) )
```

  - 16줄 : is.na() 함수에 (1, NA, 3, NA, 5)로 구성된 벡터를 전달합니다.
    - is.na() 함수는 벡터가 전달될 때 원소 하나하나에 대해 결측인지 검사하고,
    - 해당 원소와 동일한 위치에 결측이면 TRUE를, 그렇지 않으면 FALSE인 벡터를 반환합니다.
    - 첫 번째원소 1은 결측이 아니므로 FALSE, 두 번째 원소 2는 결측이므로 TRUE로 처리하고, 그 결과를 하나의 벡터로 반환합니다

```
> is.na( c(1, NA, 3, NA, 5) )
[1] FALSE  TRUE FALSE  TRUE FALSE
```

데이터 정제하기 연습

- is.na() 를 이용하여 mean() 함수에 na.rm=TRUE로 했을 때의 결과값을 구해 봅시다.

```
18. nonna.sum <- sum( ad$score[!is.na(ad$score)] )
19. nonna.length <- length( ad$score[!is.na(ad$score)] )
20. nonna.sum / nonna.length
```

  - 18줄 : ad$score에서 NA가 아닌 원소들을 추출합니다.
    - is.na() 함수가 NA이면 TRUE를 반환하므로, NA이면 FALSE가 되도록 앞에 부정을 나타내는 '!'을 붙여 ad$score에서 NA가 아닌 원소들을 추출하고, 그들의 합을 구해 변수 nonna.sum에 저장합니다.
  - 19줄 : 18줄과 마찬가지로 NA가 아닌 원소들로 추출하고, 그 원소들의 개수를 length() 함수를 이용하여 nonna.length에 저장합니다.
  - 20줄 : 결측이 아닌 자료들을 모두 합한 것을 결측이 아닌 자료들의 개수로 나누어 함수를 구합니다. 앞서 구한 mean(ad$score, na.rm=TRUE)과 동일합니다.

```
> nonna.sum / nonna.length
[1] 5.732877
```

데이터 정제하기 연습

- **factor 형 자료변환**
  - 지역규모와 성별은 1, 2, 3 등의 숫자값이 중요한 것이 아니라 1, 2, 3 각각이 나타내는 의미가 중요한 범주형 자료이나, R이 읽어 들일 때 범주형 자료인지 모릅니다.
  - 이를 위해 지역규모와 성별 변수들을 범주형 자료로 변환하기 위해 factor() 함수를 사용하여 범주형 자료로 변환하여 봅시다.
    - factor에 대한 자세한 설명은 다음 장의 "8장을 위한 준비"를 참고해 주세요.

```
22. ad$scale <- factor(ad$scale)
23. ad$sex <- factor(ad$sex)
24. str(ad)
25. summary(ad)
```

데이터 정제하기 연습

- 22, 23줄 : 함수 factor()를 이용해 기존 자료들을 범주형 자료(R 입장에서는 factor형 자료)로 변환하고, 변환된 결과로 각각 바꿉니다.
- 24줄 : factor로 잘 변환되었는지 str() 함수를 이용하여 확인합니다. 각 변수의 자료형을 나타내는 부분이 각각 3개, 2개의 수준을 갖는 factor로 변경되었음을 알 수 있습니다.
  - 수준은 범주형 자료에서 구분된 각각의 범주입니다.
  - 지역규모(scale)의 경우 대도시(특별시, 광역시), 시지역, 읍면 지역의 세 개의 범주로 되어 있어 있습니다.
  - 성별(sex)은 여성과 남성의 세 개의 범주로 되어 있습니다.

```
> str(ad)
'data.frame':  150 obs. of  4 variables:
 $ scale: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 ...
 $ sex  : Factor w/ 2 levels "1","2": 2 2 2 1 1 2 1 ...
 $ score: int  8 5 7 4 5 3 3 ...
 $ age  : int  56 33 49 53 74 42 51 ...
```

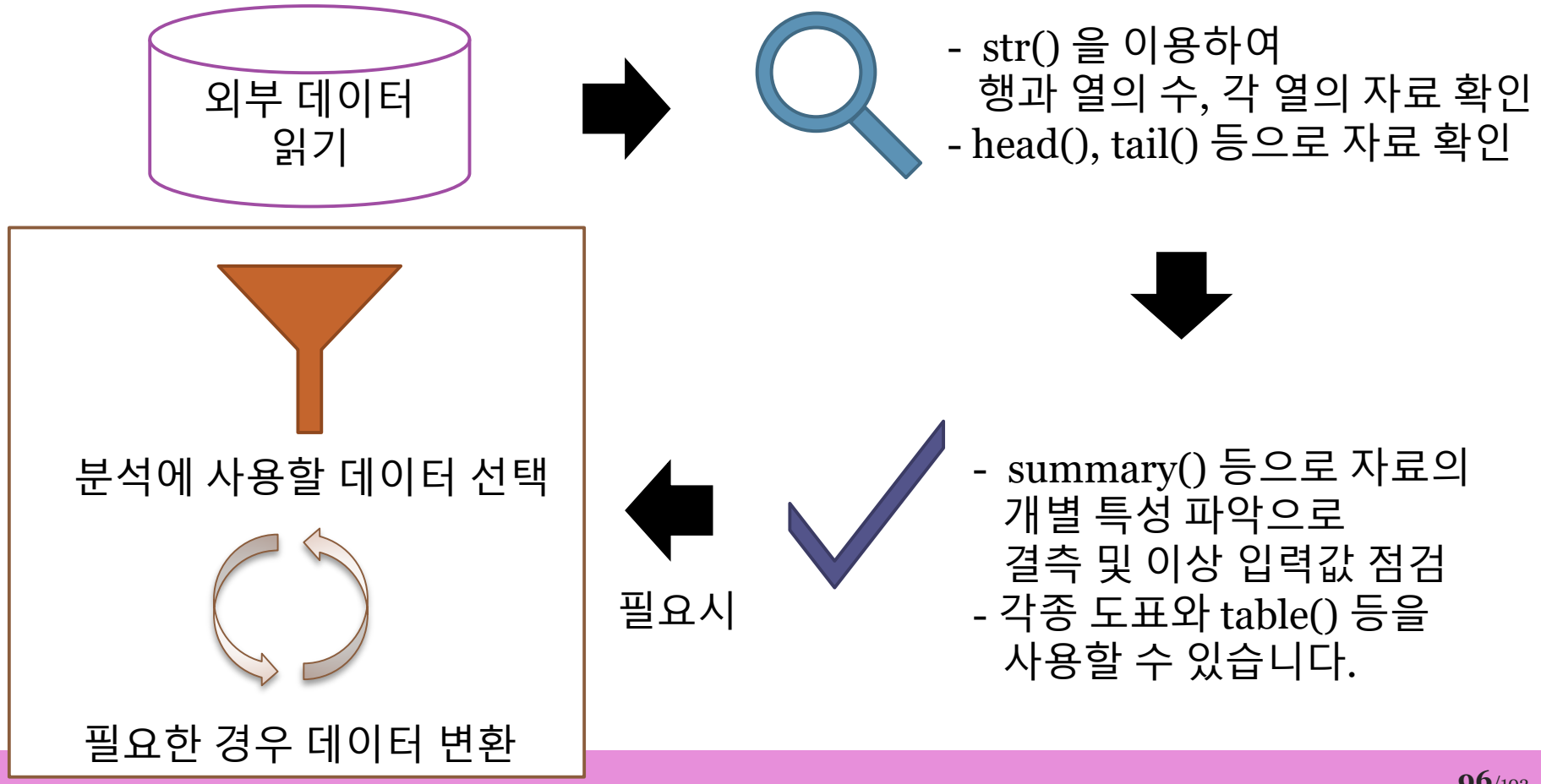
데이터 정제하기 연습

- 25줄 : summary() 함수에서 범주형(factor) 자료들을 숫자로 요약하는 방법을 확인해 봅시다. 각 수준의 수와 결측값의 수를 세어줍니다. (우리 자료에서 지역규모와 성별은 결측값이 없어 표시되지 않았습니다.)
  - 지역규모 변수를 구성하는 수준별로 각각 50개의 자료가 있습니다.
  - 성별은 여성(1)이 74명, 남성(2)이 76명으로 나타났습니다.
  - 앞서 숫자 자료일 때와 비교해 봅시다.

```
> summary(ad)
scale   sex         score             age
1:50    1:74    Min.   : 1.000   Min.   :14.00
2:50    2:76    1st Qu.: 4.000   1st Qu.:38.00
3:50            Median : 6.000   Median :46.00
                Mean   : 5.733   Mean   :46.51
                3rd Qu.: 7.000   3rd Qu.:56.00
                Max.   :10.000   Max.   :89.00
                NA's   :4
```

데이터 정제하기 연습

- **정리**
  - 다음과 같은 과정으로 분석에 사용할 자료를 준비합니다.



범주별 기초통계량 구하기

- **지역규모별 나이의 평균과 표준편차를 구하는 방법을 알아봅시다.**
  - 먼저 기존에 학습한 내용을 이용하여 범주의 값이 1이 되는 값을 선택하여 구하는 과정입니다.

```
27. length(ad$age[ad$scale=="1"])
28. mean(ad$age[ad$scale=="1"])
29. sd(ad$age[ad$scale=="1"])
```

  - 27~29줄 : ad$scale이 "1"인 자료들을 ad$age에서 추출하여 각각의 개수, 평균 그리고 표준편차를 구합니다.

```
> length(ad$age[ad$scale=="1"])
[1] 50
> mean(ad$age[ad$scale=="1"])
[1] 45.94
> sd(ad$age[ad$scale=="1"])
[1] 14.45953
```

범주별 기초통계량 구하기

- **doBy() 패키지**
  - 위의 과정을 나머지 두 범주에 대해서도 실시하는 것은 조금 번거롭습니다.
  - 각 범주별로 원하는 값을 얻기 위해 doBy() 패키지를 사용해 봅시다.
  - 먼저 다음의 두 줄로 "doBy" 패키지를 설치하고 사용할 준비를 마칩니다.
    - 인터넷에 연결되어 있어 외부 저장소에서 받는 경우입니다.

```
32. install.packages("doBy")
33. library(doBy)
```

- **summaryBy() 함수**
  - doBy 패키지에 포함된 함수로 각 범주별로 값을 사용자가 지정한 함수를 적용하여 그 결과를 보여줍니다.

범주별 기초통계량 구하기

- 기본으로 사용하는 형태는 다음과 같습니다.
  - summaryBy(formula, data = parent.frame(), FUN = mean)
  - 이외에도 많은 전달인자를 넣어 보다 다양한 작업을 할 수 있으니 help(summaryBy)를 통해 확인해 주세요.
- 기본으로 사용하는 전달인자
  - 첫 번째 전달인자(formula)로 R의 수식을 이용하여 각종 통계를 구할 변수와 집단을 구분할 변수를 표현합니다. 그 형태는 다음과 같습니다.

"통계를 구할 변수 ~ 집단을 구분할 변수"

  - data를 통해 전달되는 값은 각 변수들이 있는 데이터 프레임의 이름을 지정합니다.
  - FUN을 통해 전달되는 값은 통계를 구할 함수 이름의 벡터로 사용자 정의 함수도 사용할 수 있습니다. (값이 없을시 평균을 구합니다.)
    - 함수의 이름만 지정하며, 각 함수가 필요로 하는 전달인자는 네번째 전달인자 위치에 넣습니다.

범주별 기초통계량 구하기

- **summaryBy() 사용예**

```
> summaryBy(age~scale, data=ad, FUN=c(length))
  scale age.length
1     1         50
2     2         50
3     3         50
```

  - 데이터 프레임 ad에서 scale(지역규모)별로 age 응답의 개수(length)를 구하는 방법입니다.
    - 첫번째 전달인자로 사용한 age~scale 은 age를 scale 별로 구분함을 뜻합니다.
    - FUN을 통해 사용할 함수의 이름 length를 전달합니다. (벡터의 형태가 아니어도 실행가능하나 여러 개를 전달할 수 있음을 표현하기 위해 벡터로 나타냈습니다.)

범주별 기초통계량 구하기

- 지역규모별 나이의 평균과 표준편차를 구합니다.
  - 두 개 이상의 함수를 적용할 경우 벡터로 전달해 줍니다.
  - summaryBy() 함수에 각 함수들이 공통적으로 사용하는 전달인자를 사용합니다.

```
> summaryBy(age~scale, data=ad, FUN=c(mean, sd), na.rm=TRUE)
  scale age.mean   age.sd
1     1    45.94 14.45953
2     2    45.68 13.58937
3     3    47.92 14.87751
```

Q & A

수고하셨습니다.