

- 모집단 (population)
- 표본 (sample)



## ■ 통계학(Statistics)이란?

### ◎ 동영상: “그냥도전! 500원 동전 천 번 돌리기”

동영상 내용

- “ 500원짜리 동전을 돌렸을 때 학이 나올 확률이 70% 정도
- 500원짜리 동전을 1000번 돌리는 실험
- 실험 결과 1000번 중 학이 679번이 나옴
- 학이 나올 가능성이 70%정도 된다는 것이 얼추 맞다는 주장

## ◎ 통계학적 관점 (614)

- ① ● 500원짜리 동전을 돌렸을 때 학이 나올 확률이 70% 정도  
⇒ 관심 또는 연구의 대상(문제, 주제, 가설) *가설*
- ② ● 500원짜리 동전을 1000번 돌리는 실험  
⇒ 실험을 통해 자료를 수집
- ③ ● 실험 결과 1000번 중 학이 679번이 나옴  
⇒ 수집된 자료를 정리, 요약, 분석하여 자료의 특성을 파악

- ④ ● 학이 나올 가능성이 70%정도 된다는 것이 얼추 맞다는 주장  
⇒ 자료의 특성을 이용하여 관심 또는 연구의 대상에 대해  
추론

## ◎ 관심의 대상 vs 자료

500원 짜리 동전의 H 횟수

1000번 반복 실험 결과

"모집단"

④

"표본"

## ◎ 모집단(population)

- 잘 정의된 연구목적과 이와 연계된 명확한 연구대상을 설정

예] 대통령 후보의 지지율? ⇒ 유권자 목표에 맞는 연구 대상.

- **연구대상이 되는 모든 개체의 집합**

예] 19대 대통령선거 선거인명부 유권자수는 42,432,413명

Q. "그냥도전!"에서의 모집단은? 무한히 많은 경우.

∞

## ◎ 전수조사: 모집단 전체를 대상으로 조사하는 경우

- 조선(대한제국포함)시대 임금의 수명
  - 27명의 임금의 수명 자료
- 2010년까지의 인구주택총조사(census)

## ◎ 대부분의 모집단은 매우 커 전체를 조사하기 어려움

- 적절한 방법으로 일부의 자료를 추출해 조사
- ex) ● 2015년부터 행정자료 + 20% 표본조사

del) "표본" (sample)

: 모집단으로부터 작은 일부의 개체를 뽑아서 조사.

## ◎ 표본(Sample)

- 모집단으로부터 선택된 일부의 개체  
예】 “그냥 도전!” 에서 나온 1,000번의 동전 결과  
예】 각종 여론조사에 참여한 유권자

표본 추출시 고려사항

Q1. 추출된 표본이 모집단을 대표할 수 있는가?

Q2. 몇 명(개)의 표본을 어떻게 뽑아야 하는가?

↑ “통계학의 작은 part.”

08)

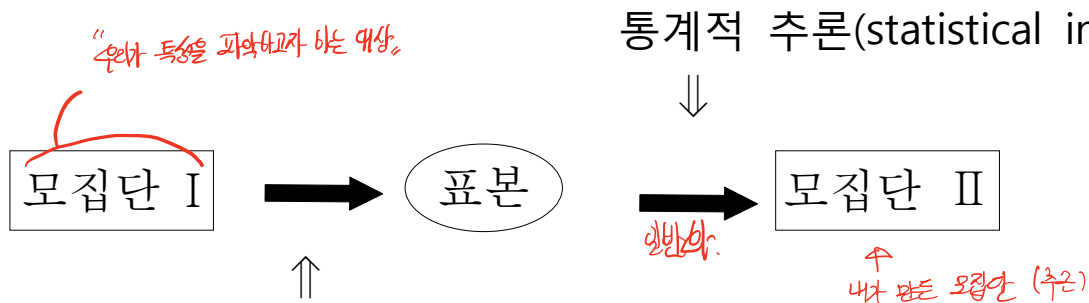
## ◎ 1936년 미국대통령 선거

- 공화당의 Landon vs 민주당의 Roosevelt
- ‘Literary Digest’
  - 1916~1932년 선거결과 정확하게 예측
  - 구독자, 전화기 및 자동차 보유자 236만여 명의 의견을 분석  
⇒ Landon 57%, Roosevelt 43%
- ‘Gallup’ : 5만 명의 표본조사(할당추출)  
⇒ Landon 44%, Roosevelt 56%

표본추출 방식에 따라  
다른 결과를 도출.

- 선거결과: Roosevelt 63%, Landon 37%
  - 실제 득표율과 7% 차이를 보임
- 1948년 Dewey vs Truman 대결에서 Gallup 승자 예측 실패  
⇒ 표본선정방식 변경: 할당추출법 ⇒ 확률추출법

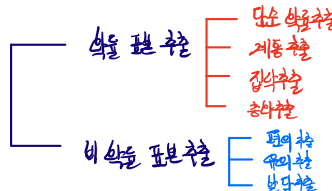




- (설문)조사(survey)
- 실험(experiment)
- 관찰(observation)
- $\vdots$
- 전수조사, 빅데이터(?): 모집단  $\approx$  표본

## ◎ 통계학이란

- 관심 또는 연구의 대상인 모집단의 특성을 파악하기 위해
- 모집단부터 일부의 자료(표본)를 수집하고
- 수집된 표본을 정리, 요약, 분석하여 표본의 특성을 파악한 후
- 표본의 특성을 이용하여 모집단의 특성에 대해 추론하는  
원리와 방법을 제공하는 학문



## ◎ 확률표본추출 vs. 비확률표본추출

### ● 확률표본추출(probability sampling)

- 모집단을 구성하는 모든 추출단위에 대해 표본으로 추출된 확률을 알 수 있는 추출법 즉, 특정 추출에 기반한 선택의 방식.

예를 들어, ⇒ 표본추출틀(sampling frame, 표집틀) 필요

- 예】 모집단: {①, ②, ③, ④, ⑤} ⇒ 2개의 표본

· 어떤 개체가 표본으로 뽑힐 확률 =  $2/5$

- 특정한 표본이 선정될 확률을 토대로 추정오차를 확률개념을 이용하여 과학적으로 설명

대표적 종류 ○ 단순확률추출, 계통추출, 집락추출, 층화추출 등

- **비확률표본추출(non-probability sampling)**

- 특정 표본이 선정될 확률을 알 수 없음

⇒ 추론결과의 정확도(precision)? X *추론에 대한 정확도를 알 수 없다. → 즉, 표본의 특성을 모집단의 특성과 유사하게 하기 위해 존재한다.*

- 예】 편의(convenience)추출, 유의(purposive)추출, 할당(quota)추출

- **편의추출**: 자발적 참여, 백화점 앞, 포털사이트 인터넷 조사 *(조사자의 편의에 기함)*

- **유의추출**: 전문가 선택 *(연구자 대상으로)*

- **할당추출**: 그룹 내 조사대상 선택에서 **랜덤화 과정 없음** < *상대별-구분, 연령별-구분, ...*

- 간편하고 비용이 적게 든다는 이유로 사회조사에서 광범위하게 사용됨

\* 표집의 구분

## ◎ 목표모집단 vs 조사모집단

- 목표모집단(target population)
  - 관심대상이 되는 모든 기본단위들의 집합 (일반적으로 원하는 표집과 동일)
  - 시공간상 명확하게 정의된 연구대상 집단
    - 조사시점, 지리적인 경계, 연령 기준 등
    - 예】 수도권 거주 고등학생 학부모 대상 조사
      - ↳ 표본추출틀 필요
      - 하지만 구상하기 애매함

- 조사모집단(survey population)
  - ④ 조사가능모집단(accessible population)
    - (현실적인 제약 고려) 표본추출 대상 기본단위들의 집합
    - 표본추출틀(sampling frame)을 통해 추출될 수 있는 기본단위들의 집합
      - 예】 전화여론조사: 전화번호부(표본추출틀)에 등재된 전화보유 가구의 성인

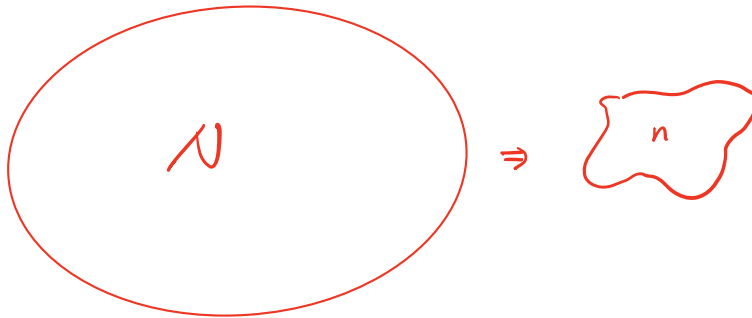
∴ 실제 조사에서는 → 최초 목표로 설정한 '목표 모집단'과 실제 대상인 '조사 모집단'을  
인식을 필요로 한다.

## ◎ 확률표본추출방법의 종류

- 단순확률추출법 (Simple random sampling) - 무작위 추출
- 계통추출법 (Systematic sampling)
- 층화확률추출법 (Stratified random sampling)
- 집락추출 (Cluster sampling)

## ① 단순확률추출 (SRS, simple random sampling)

- 크기가  $N$ 인 모집단에서 크기  $n$ 인 표본을 무작위로 추출
- 모든 단위들이 표본에 선택될 확률이 동일
  - 예】 가구조사:  $P(\text{이니네 집 추출}) = n/N$
- 실제 대규모 조사에서는 거의 사용되지 않지만 다른 모든 표본추출방법의 기초 (표본추출 기초 이론)

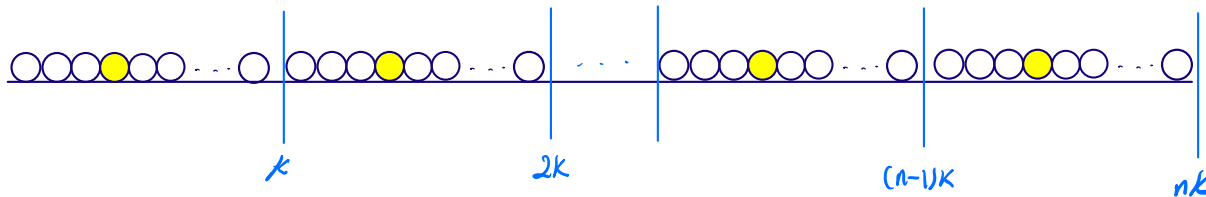




## ② 계통표본추출(systematic sampling)

- 표집틀에서 처음  $1 \sim k$  번째 단위들 중 하나를 랜덤하게 선택한 다음, 매  $k$  번째에 해당되는 단위들을 표본으로 추출

- 계통표본 추출과정
  - 500개 중 10개 추출  $\Rightarrow 500/10$   $\rightarrow 3, 50+3, \dots, 450+3$
  - 추출간격  $k$  의 결정:  $N/n$  또는 정확도를 고려 결정
  - $1 \sim k$  에서 난수 하나를 선택해서 시작점을 선정
  - 시작점에  $k$  를 반복적으로 더해서 표본추출



◎ 모집단 크기 = 500, 표본크기 = 50



1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
491	492	493	494	495	496	497	498	499	500

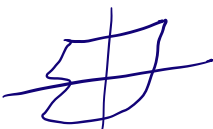
↪ 표본추출:  $\frac{1}{10}$ 
↪ 표본추출:  $\frac{1}{10}$ 
↪ 표본추출:  $\frac{1}{10}$

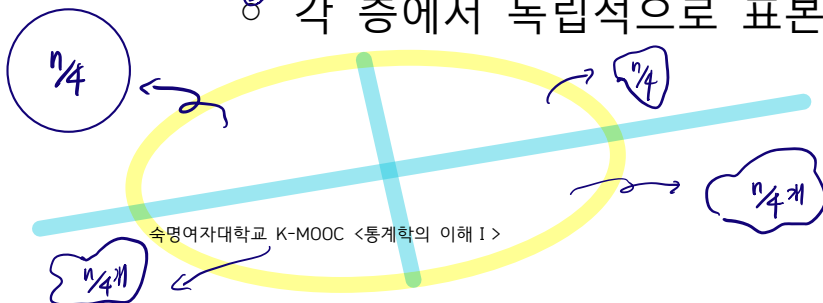
- 표집틀이 없어 고유번호 부여, 난수발생 등 단순확률추출법을 적용하기 어려운 실제 조사현장에서 폭 넓게 활용
  - 예】 선거출구조사, 주차장 출입 차량에 대한 조사



### ③ 층화확률추출(stratified random sampling)

왜냐고 가장 많은

- 모집단을 서로 중복되지 않는 여러 개의 층(strata)으로 나누고, 각 층에서 단순확률추출에 의해 표본을 추출
  - 부모집단(subpopulation)의 구성 내역을 알고 있음
  - 부모집단 간 특성에 차이가 있음 즉, 모집단을 나눌 수 있는 기준 있어야 함.
  - 전체 모집단 크기  $N$ ,  $i$  번째 층의 크기  $N_i$ ,  $W_i = N_i/N$
- 층화 표본추출 과정
  - ① 층의 구성(성별, 연령, 지역 등) 
  - ② 각 층에서 독립적으로 표본 추출  $\Rightarrow$  단순확률추출 사용



- ◎ 서울시내 서점의 월 매출액 추정을 500개 서점 표본추출
    - 500개 중 대형 서점이 10개인 경우와 20개인 경우
      - ⇒ 추정치 변동이 큼
    - 대형, 중형, 소형으로 분류 후 각 층에서 일정 수 표본추출
      - 층의 비율에 맞게 추출 *OK!*
      - 층의 비율에 맞지 않으면 가중치 반영
- (대형 - 3개 점  
중형 - 10개 점  
소형 - 50개 점) ⇒ 가중치 부여*

#### ④ 집락표본추출 (cluster sampling)

- 서로 인접한 조사단위들을 묶어 구성한 집락(cluster)을  
추출하고, 이들 집락 내의 조사단위들을 조사
- 예】 서울시 고등학생 월평균 사교육비 추정



## □ 요약

- 확률표본추출 vs 비확률표본추출
- 확률표본추출
  - 단순확률추출법
  - 계통추출법
  - 층화추출법
  - 집락추출법

\* 선거에 대한 개표방식. — 관할 및 관할권 외장 중 → 관할권 구제 다른 표의 처리는 정 → 개표 과정을 통한 해결

[1주] 통계학이란? - 3. 가중치

9x

## ◎ 개표방식

○ 지역구: A지역 7만 명 투표, B지역 3만 명 투표 — 총 10만

○ 개표율: A지역 10%, B지역 50% —  $\frac{1}{10}$   $\frac{1}{2}$   $\frac{1}{10}$   $\frac{1}{2}$

· A지역 ①번 후보자 득표율 60%, ②번 후보자 40%

· B지역 ①번 후보자 득표율 30%, ②번 후보자 70%

⇒ ①번 후보자 득표수:  $\frac{7\text{만} \times 0.1 \times 0.6}{\text{A지역 총 표수} \quad \text{개표율} \quad \text{득표율}} + \frac{3\text{만} \times 0.5 \times 0.3}{\text{B지역 총 표수} \quad \text{개표율} \quad \text{득표율}} = 0.87\text{만명}$

②번 후보자 득표수:  $\frac{7\text{만} \times 0.1 \times 0.4}{\text{A지역 총 표수} \quad \text{개표율} \quad \text{득표율}} + \frac{3\text{만} \times 0.5 \times 0.7}{\text{B지역 총 표수} \quad \text{개표율} \quad \text{득표율}} = 1.33\text{만명}$

⇒ ① 득표율:  $\frac{0.87}{(0.87+1.33)} = 39.5\%$  ② 득표율: 60.5%

issue · (개표된 A지역의 한 표는 10표, B지역의 한 표는 2표를 대표) ← A지역 개표 10%, B지역 개표 50%

- 해당지역의 득표율이 유지된다면
  - ①번 후보자 득표수:  $7\text{만} \times 0.6 + 3\text{만} \times 0.3 = 5.1\text{만} \Rightarrow 51.0\%$
  - ②번 후보자 득표수:  $10\text{만} - 5.1\text{만} = 4.9\text{만} \Rightarrow 49.0\%$



## ◎ 가중치(weight)

- 모집단의 구성정보는 표본을 추출하는데 있어 매우 중요한 사전정보
    - ⇒ 표본조사 결과의 정확도를 높일 수 있는 핵심요소
  - 모집단이 서로 다른 특성을 가지는 부모집단들로 이루어진 경우, 특정 부모집단에서 표본이 많이 추출되거나 적게 추출되면 전체 모집단에 대해 왜곡된 결과가 나올 수 있음
    - 표본추출설계에 충실히 반영해도 실제 표본획득 과정에서 문제가 발생
      - ⇒ 가중치 적용
- basic concept* 한 표본이 몇 개를 대표하는지

- 기본 가중치

- 단순확률추출법: 각 표본에 대한 설계가중치:  $w_j = N/n$  예) 500개 중 10개 추출 → 1개의 객체가 각각 50개를 대표
- 계통추출법: 각 표본에 대한 설계가중치:  $w_j = N/n = k$  → 각 실을 위해 1차원 74명 집안으로
- 층화확률추출법:
  - 층의 크기와 해당 층에서의 표본크기에 따라 달라짐
- 집락추출:
  - 집락의 크기와 해당 집락에서의 표본크기에 따라 달라짐

# \* 층화 추출 방식에서의 가중치 부여 방법

[1주] 통계학이란? - 3. 가중치



- ① 추출확률에 따른 가중치:  $w_1$
- ② 무응답에 따른 가중치:  $w_2$
- ③ 사후층화를 위한 가중치:  $w_3$

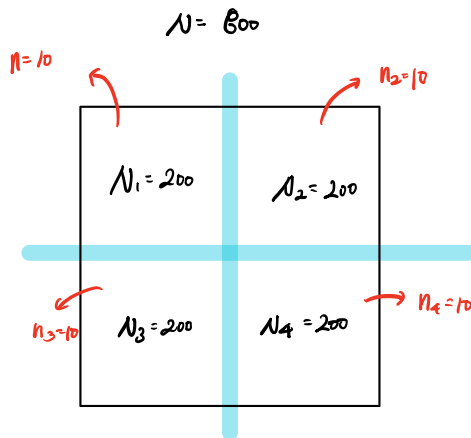


SOOKMYUNG  
WOMEN'S UNIVERSITY  
K-MOOC

## ① 추출확률에 따른 가중치: $w_1$

### ● 등확률 추출인 경우

- 표본으로 선택될 확률 =  $n/N$
- 표본에서 차지하는 비중 =  $1/n$   
 $\Rightarrow$  표본 한 명이  $N/n$  명을 대표



### ● 등확률 추출이 아닌 경우

- 추출확률의 상이함에 따른 조정
- 설계 가중치, 표본추출 가중치, 기초 가중치



ex) ~~설계~~ × ~~해결~~

## ◎ 대학졸업자 취업 현황조사

○ 수도권과 지방 대학 2014년 총

○ 모집단: 수도권 = 40만, 지방 = 20만

○ 표본크기: 수도권 = 500, 지방 = 500

○ 추출률: 수도권 = 5백/40만 = 1/800, 지방 = 1/400

○ 설계가중치 = 1/추출률

· 수도권 표본 한명이 800명을 대표

· 지방 표본 한명은 400명을 대표

## ② 무응답에 따른 가중치: $w_2$

조사 대상 표본이 응답 시,  
이들 대채 시가 위한

24 대상 표본이

- 대체표본이 없거나 일부 항목에 답을 하지 않은 경우

### ◎ 대학졸업자 취업 현황조사

- 응답률: 수도권 = 60%, 지방 = 80% 응답
- 응답가중치 = 1/응답률

· 수도권 응답자의 응답가중치 =  $10/6$

⇒ 수도권 응답자 1인당  $800 \times 10/6 = 1333.3$ 명 대표

· 지방 응답자의 응답가중치 =  $10/8$

⇒ 지방 응답자 1인당  $400 \times 10/8 = 500$ 명 대표

· 조사 대상, 표본에 대한 사전 정보(특성)로 분류

→ 설문조사에 응답자의 성별, 연령대 등을 고려

→ 사후(설문 후) 응답을 다시 가하는 것

⇒ '사후 응답'에 대한 가중치



### ③ 사후증화를 위한 가중치: $w_3$

- 가중 표본 분포가 어떤 특성에 대해 알려진 모집단 분포와 일치하도록 조정

### ● 대학졸업자 취업 현황조사 ① 지역별 응답률

- 성별에 따라 취업 현황에 차이가 있음(가정) - ② 새로운 특성 발견
- ③ 수도권과 지방 졸업자의 성별 구성은 비슷함(가정)
- 남녀 비율: (45%, 55%), 표본에서의 비율: (60%, 40%)
- 사후증화 가중치 (모집단)

$$\Rightarrow \left( \begin{array}{l} \cdot \text{남자의 가중치} = \frac{45}{60} \\ \cdot \text{여자의 가중치} = \frac{55}{40} \end{array} \right)$$

- 최종 가중치(final weight):  $w_f = w_1 \times w_2 \times w_3$ 
  - 응답한 수도권 남자:  $w = 800 \times \frac{10}{6} \times \frac{45}{60} = 1000$
  - 응답한 수도권 여자:  $w = 800 \times \frac{10}{6} \times \frac{55}{40} = 1833.3$
  - 응답한 지방 남자:  $w = 400 \times \frac{10}{8} \times \frac{45}{60} = 375$
  - 응답한 지방 여자:  $w = 400 \times \frac{10}{8} \times \frac{55}{40} = 687.5$

□ **요약:**  $w_f = w_1 \times w_2 \times w_3$

- $w_1$ : 확률추출에 따른 가중치
- $w_2$ : 무응답에 따른 가중치
- $w_3$ : 사후층화를 위한 가중치

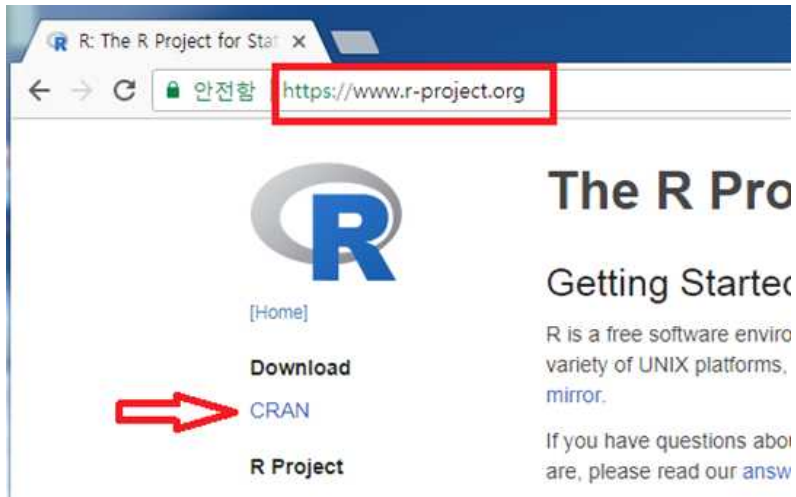


## ■ 통계프로그램

- 상용프로그램
  - 상업적 목적이나 판매 목적으로 만든 프로그램
  - SAS, SPSS, STATA, Matlab
- 무료프로그램
  - **R**, Python ⇒ 일종의 고급 컴퓨터 언어
  - SAS University Edition ⇒ 가상 애플리케이션

- R 설치

- <http://www.r-project.org> 접속



- CRAN Mirrors에서 "Korea"에 있는 사이트 중 하나 선택
  - <http://healthstat.snu.ac.kr/CRAN>
  - <https://cran.biodisk.org>
  - <http://cran.biodisk.org>
- OS에 맞는 버전 선택

Download and Install R

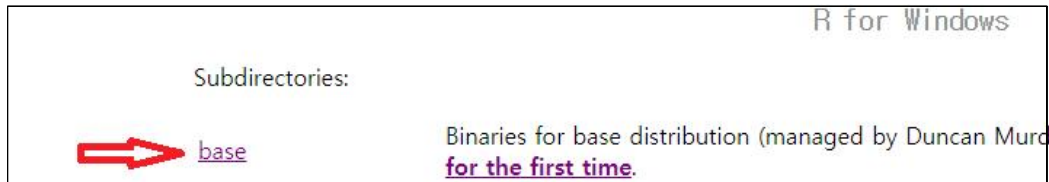
Precompiled binary distributions of the base system and most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)



※ OS를 Window로 선택했을 경우입니다.

- “base” 클릭



- “Download R 3.5.0 for windows” 클릭



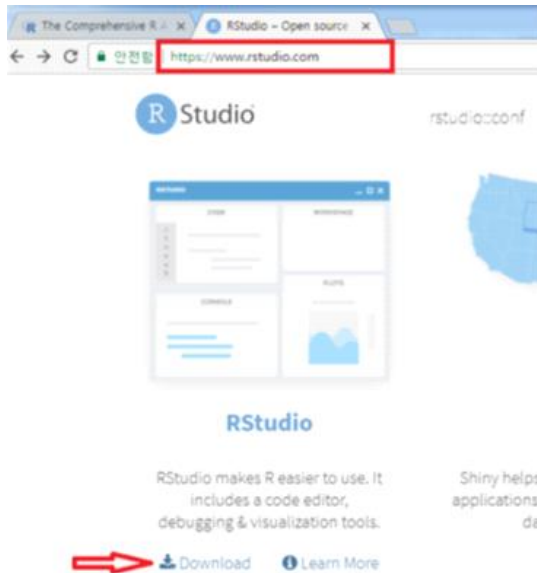
※ 프로그램이 최신 버전으로 업데이트되었을 경우 버전이 설명 이미지와 다를 수 있습니다.

- 설치 프로그램 실행

※ 설치 시 32bit와 64bit가 구분됩니다. 윈도우 환경을 확인하시고 설치하시기 바랍니다.

## • R Studio 설치

- <http://www.rstudio.com> 접속 > Download 클릭



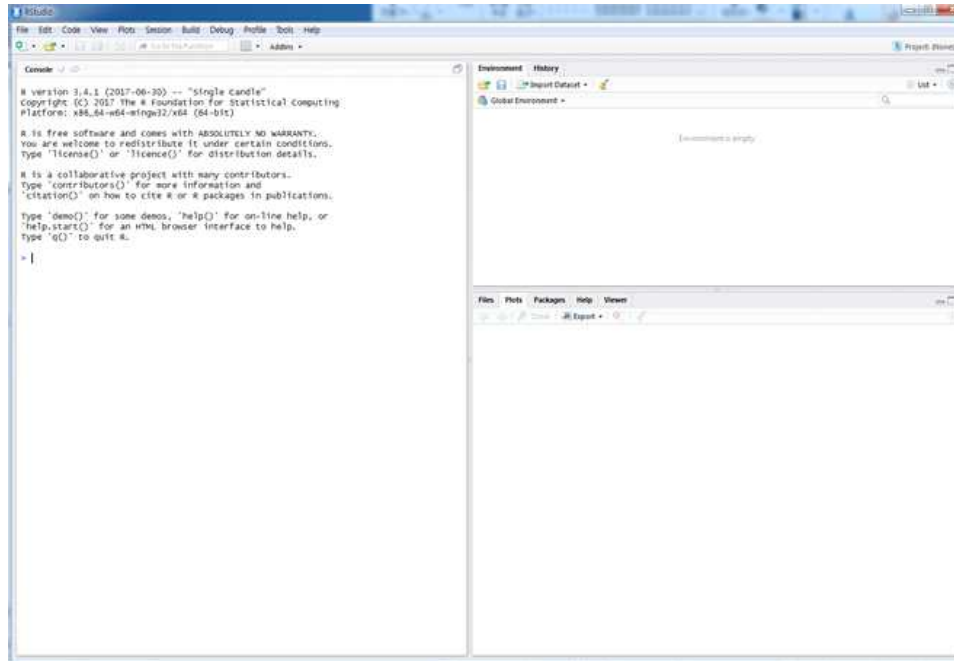


- OS에 맞는 Rstudio 선택

Installers for Supported Platforms		
Installers	Size	
RStudio 1.0.153 - Windows Vista/7/8/10	81.9 MB	
RStudio 1.0.153 - Mac OSX 10.6+ (64-bit)	71.2 MB	
RStudio 1.0.153 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	85.5 MB	
RStudio 1.0.153 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	91.7 MB	
RStudio 1.0.153 - Ubuntu 16.04+/Debian 9+ (64-bit)	61.9 MB	
RStudio 1.0.153 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	84.7 MB	
RStudio 1.0.153 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	85.7 MB	

- 설치 프로그램 실행

## ● R Studio 실행





## ■ 과제 1

- 국가통계포털(<http://kosis.kr>)에서는 국가승인통계 제공
- 실업률과 관련된 통계지표를 찾아 통계명, 목표모집단, 조사모집단, 표본추출법, 표본수, 조사주기에 대해 알아보기

## ■ 과제 2

표본	구분	지역1	지역2	지역3	합
	50대 이상	150	200	150	500
	40대 이하	150	100	250	500

모집단	구분	지역1	지역2	지역3	합
	50대 이상	4,000	3,000	3,500	10,500
	40대 이하	5,500	5,000	4,500	15,000

- 지역의 정보만을 이용하여 지역 1의 표본에 대한 가중치를 유도하기
- 지역과 연령정보를 이용하여 지역 1의 50대 이상의 표본에 대한 가중치를 유도하기