표본 추출(Data Sampling)

표본추출 종류

표본추출 방법	
단순 임의 추출	전체 데이터에서 각 데이터를 추출할 확률을 동일하게 하여 표본을 추출하는 방법
가중치를 고려한 표본 추출	각 데이터의 중요도나 발생 빈도가 다르다면 이를 고려한 표본추출 방법
층화 임의 추출	모집단을 보다 동질적인 몇 개의 층(Strata)으로 나눈 후, 이러한 각 층으로부터 단순 임의 표본추출을 하는 방법
계통 추출	모집단 목록에서 구성요소에 대해 일정한 순서에 따라 매 K번째 요소를 추출하는 방법

1. R에서 단순 임의 추출 실시하기

- 단순 임의 추출하는 방법

- 복원 추출(Sampling with Replacement) : 한 번 추출된 표본을 다시 선택하는 것이 가능

- 비복원 추출(Sampling without Replacement) : 비복원 추출은 한 번 추출한 표본은 다시 추출하지 않는 것이 불가능

- R함수 : sample( )

```
> sample(1:10, 5) #비복원 추출
[1]  6  9  7  1 10
> sample(1:10, 5, replace = TRUE) #복원 추출
[1]  1  9  1  4 10
```

2. R에서 가중치를 고려한 표본 추출

- 단순 임의 추출하는 방법
- 각 데이터의 중요도나 발생 빈도가 다르다면 이를 고려하여 표본을 추출해야 함
- 이 경우 각각의 모집단의 원소에 다르게 가중치를 두어 표본추출될 확률을 다르게 할수 있음
- 가중치를 고려한 표본추출은 sample() 함수에서 prob를 원소별로 다르게 두어 수행
- R함수 : sample(x, size, replace=TRUE or FALSE, prob = 각표본이 뽑힐 확률)

```
> sample(1:10, 5, replace = TRUE, prob = 1:10) #가중치를 고려한 표본추출
[1] 6 7 8 5 6
```

- - 숫자가 높을 수록 높은 가중치를 두었기 때문에 높은 숫자가 더 많이 추출됨을 확인

3. R에서 층화 임의 추출 실시하기

- 모집단의 데이터가 중첩 없이 분할 될 수 있는 경우, 그리고 각 분할의 성격이 명확이 다른 경우, 층화 임의 추출을 수행하여 더 정확한 분석 결과를 얻을 수 있음
- R함수: sampling( )패키지의 strata()함수

sampling 패키지의 strata() 함수의 입력인자	
data	데이터를 추출할 데이터 프레임 또는 행렬
strataname	층화 추출에서 사용할 변수들
size	각 층의 크기
method	데이터를 추출할 방법 ("srswor" : 비복원 단순 임의 추출 "srswr" : 복원 단순 임의 추출 "poisson" : 포아송 추출 "systematic" : 계통 추출)
pik	각 데이터를 표본에 포함할 확률
description	TRUE이면 표본의 크기와 모집단의 크기를 출력함

3. R에서 층화 임의 추출 실시하기

- strata() 함수의 출력은 데이터 값이 아닌 rownum 기준임으로 추출된 표본의 변수들을 확인할려면 sampling 패키지의 getdata() 함수를 이용

sampling 패키지의 getdata() 함수의 입력인자	
data	데이터를 추출할 데이터 프레임 또는 행렬
m	선텍된 행에 대한 벡터 또는 표본 데이터 프레임

```
install.packages("sampling")
library(sampling)
x <- strata(c("Species"), size = c(3, 3, 3), method = "srswor", data=iris) #층화임의 추출
x
getdata(iris, x)
iris[x$ID_unit,]
```

4. R에서 계통 추출 실시하기

- 계통 추출은 모집단의 임의 위치에서 시작해 k번째 항목을 표본으로 추출하는 방법으로 모집단이 어떤 특징을 갖고 있을 때, 단순임의추출법보다 계통추출법의 정도가 더 좋음
- R함수 : 계통 추출은 doBy 패키지의 sampleBy() 함수

doBy패키지의 sampleBy() 함수의 입력인자	
formula	~ 우축에 나열한 이름에 따라 데이터가 그룹으로 묶음
frac = 0.1	추출할 샘플 비율 기본값은 10%
replace	복원 추출 여부
data = parent.frame()	데이터를 추출할 데이터 프레임
systematic = FALSE	계통 추출(Systematic Sampling)을 사용할지 여부

4. R에서 계통 추출 실시하기

- 다음 예는 1:10을 저장한 데이터 프레임에서 3개의 표본을 계통 추출로 뽑는 예
- 코드에서 sampleBy() 함수의 첫번째 인자는 '~1'인데 그 이유는 첫 번째 인자가 표본을 추출할 그룹을 지정하는 formula이기 때문임. 여기에서는 그룹의 구분이 필요 없으므로 상수 1을 사용

```
install.packages("doBy")
library(doBy)
x <- data.frame(x = 1:10)
sampleBy(~1, frac = .3, data=x, systematic = TRUE)
```

5. Train data set, Test data set 구분

- Train Data = 모델의 훈련을 위한 **훈련용 데이터**
- Test Data = 모델을 평가하기 위해 정답(결과)을 이미 알고있는 **테스트용 데이터**

(1) Sample() 함수 활용

(2) Row 번호 활당을 통한 샘플링

(3) sampleBy() 함수를 활용

(4) caret::createDataPartition() 함수를 사용한 샘플링