

데이터 전처리 방법

데이터 구조의 정량, 정성적 파악

□ 데이터 구조의 정량적/정성적 파악

- 데이터에 대한 전반적인 이해를 통해 분석 가능한 데이터인지 확인하는 단계
- 분석이 불가능한 상태라면 분석 가능한 데이터로 만들기 위한 방향을 설정하기 위해 필요

□ 분석 가능한 데이터인지 확인하기 위한 항목

- 변수 유형 확인 : num, char, factor 등 변수의 유형 확인
- 표준화의 필요성 확인 : 변수 별로 다양한 단위가 존재할 경우, 변수의 표준화와 필요 여부 확인
- 변수 제거 여부 결정 : 변수 별 결측률 확인 & 대체값 적용 시 자료의 변화를 확인한 후 변수제거 여부 결정
- 분석에 필요한 데이터 양 확인 : 독립변수 대비 충분한 데이터량이 확보되는지 확인

데이터 구조의 정량, 정성적 파악

□ 변수 유형 확인

- Character : 문자형 (예 : SB66885)
- Numeric : 숫자형 (예 : 34.6)
- Factor : 범주형 (예 : M/F)
- Integer : 정수형 (예 : 5)

예를 들어,

- 속도라는 숫자형 데이터에 “M”과 같은 변수의 유형에 맞지 않은 데이터가 포함되어 있을 경우를 확인하여 수정하여야 한다.

데이터 구조의 정량, 정성적 파악

□ 표준화의 필요성

- 예를 들어 두 개의 데이터를 비교하고자 하는 경우, 두 데이터의 단위가 너무 다르기 때문에 비교하기 힘든 경우가 있다. 이런 경우, 원활한 비교를 위해 데이터를 표준화할 필요가 있다.

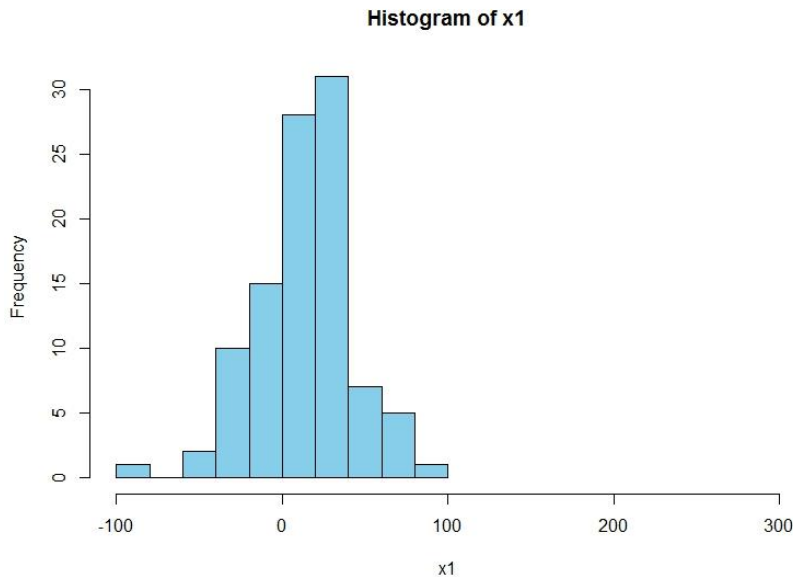
□ 표준화 방법

- $Z_i = \frac{X_i - \bar{X}}{s}$
- Z_i : 표준화한 데이터
- X_i : 원 데이터(original values)
- $\bar{X}$: 데이터의 평균
- s : 데이터의 표준 편차

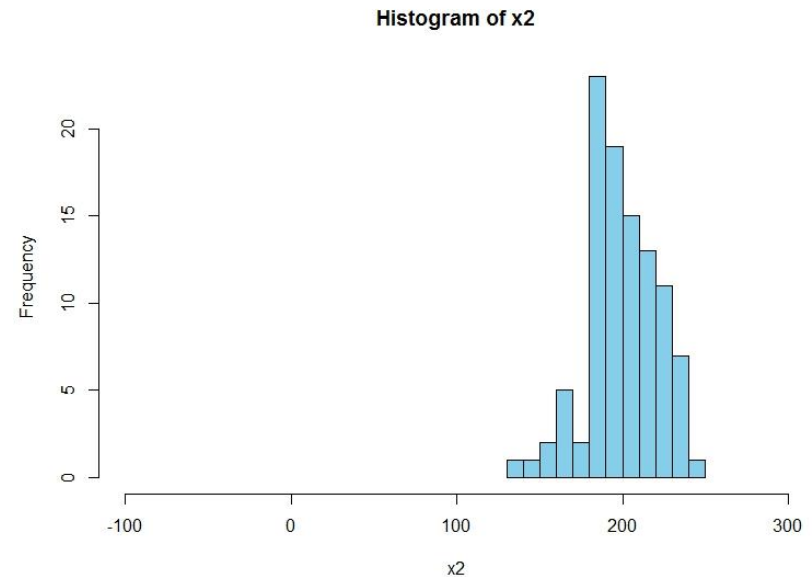
데이터 구조의 정량, 정성적 파악

□ 표준화의 필요성

- 표준화 전 두 Sample data의 분포 (histogram)



Mean=10
Standard deviation = 30

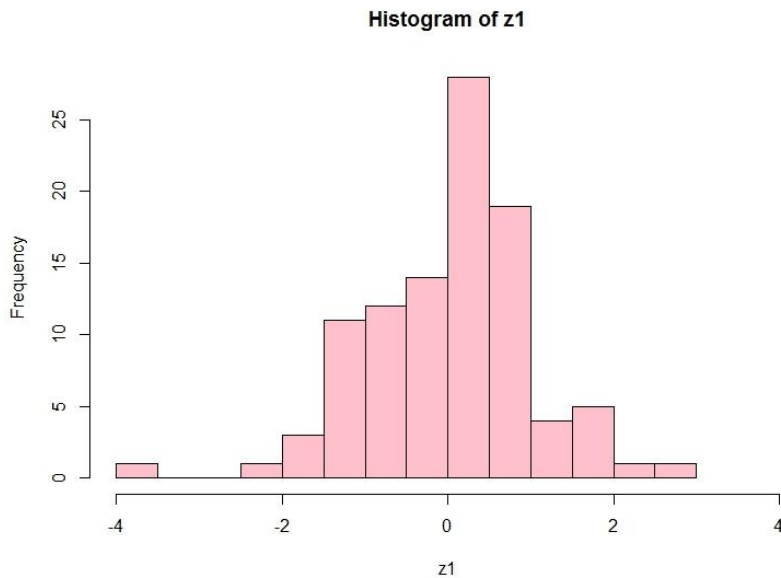


Mean=200
Standard deviation = 20

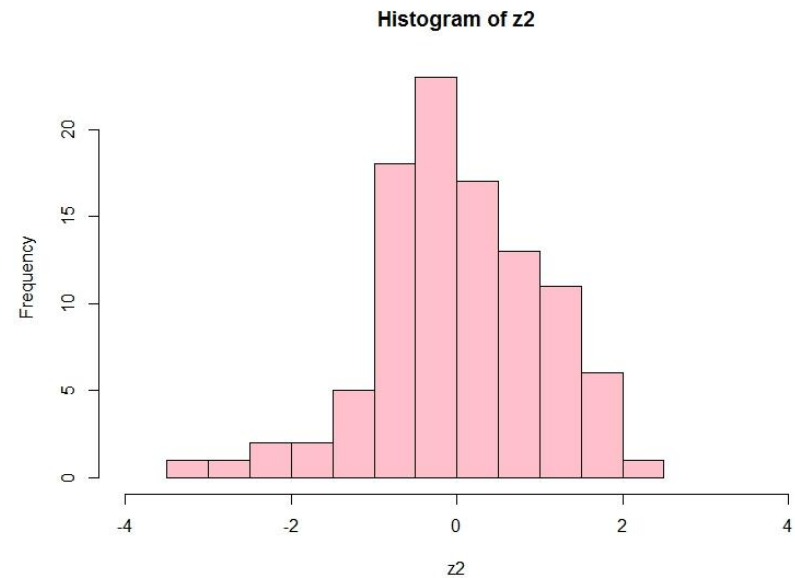
데이터 구조의 정량, 정성적 파악

□ 표준화의 필요성

- 표준화 후 두 Sample data의 분포 (histogram)



Mean=0
Standard deviation = 1



Mean=0
Standard deviation = 1

데이터 구조의 정량, 정성적 파악

□ R 예제코드

```
# Original values
set.seed(1234)
x1 <- rnorm(100, mean=10, sd=30)
hist(x1, xlim=c(-100,300), breaks=10, col="skyblue")

x2 <- rnorm(100, mean=200, sd=20)
hist(x2, xlim=c(-100,300), breaks=10, col="skyblue")

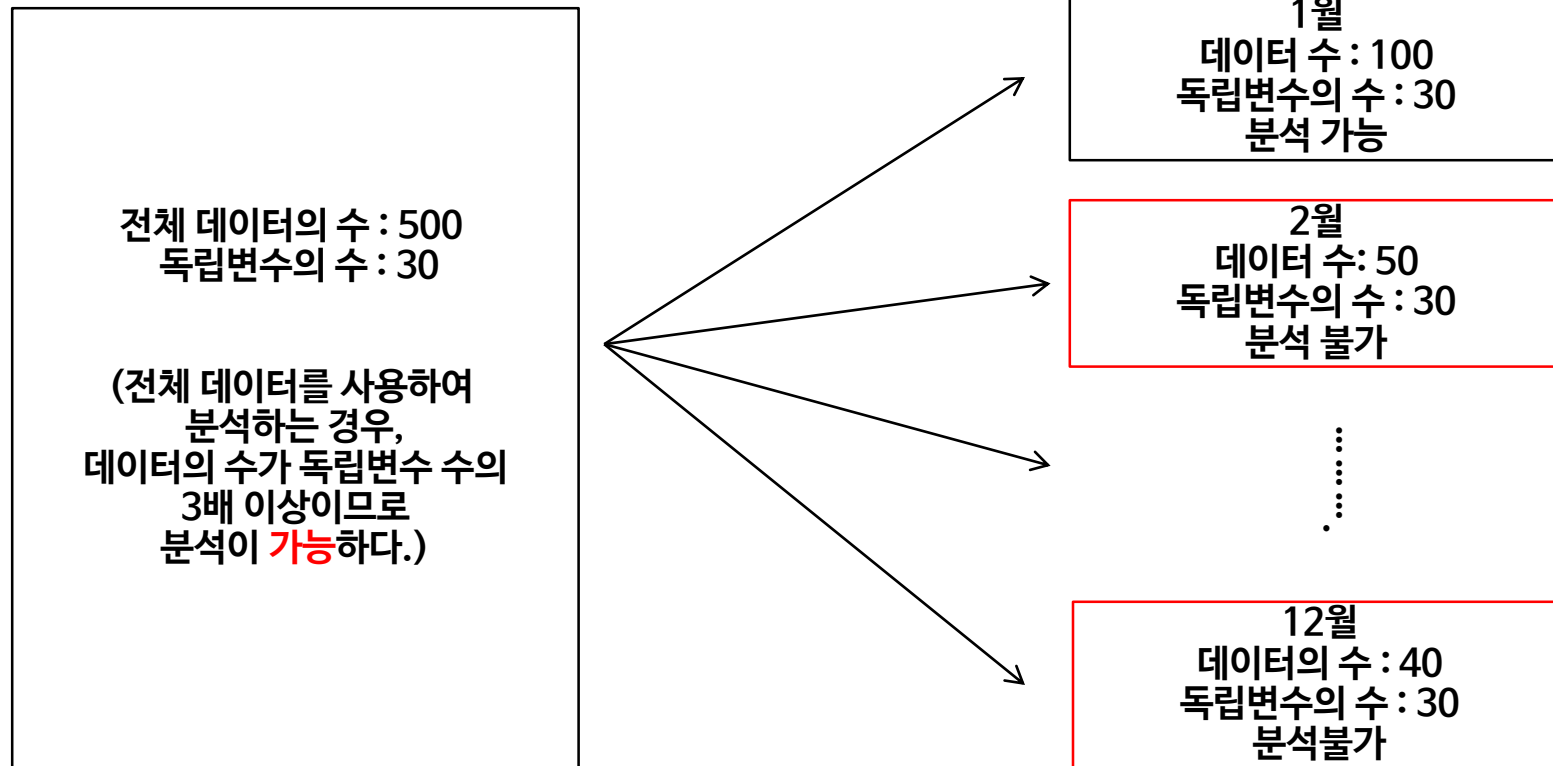
# Standardize values
z1 <- (x1-mean(x1))/sd(x1)
hist(z1, xlim=c(-4,4), breaks=10, col="pink")

z2 <- (x2-mean(x2))/sd(x2)
hist(z2, xlim=c(-4,4), breaks=10, col="pink")
```

데이터 구조의 정량, 정성적 파악

□ 데이터 양 확인

- 독립변수 수의 3배 이상 정도가 되어야 한다.



이상값, 결측값, 분포해석 및 정제 방법

□ 이상값

- 변수의 분포에서 비정상적으로 벗어난 값으로 Box-plot을 통해 분포 및 이상값을 확인 할 수 있다.

□ 결측값

- 값이 관측되지 않은 자료
- 해당 칸이 비어져 있는 경우 알기 쉬우며 보기에는 값이 관측된 듯 해당되는 관측값이 있지만 실상 Default값이 기록된 경우도 있다.

□ 분포 분석

- 막대그래프, Histogram, Box-plot, 기술통계량 등을 통하여 분포를 파악한다. 정규분포가 아닌 경우 변환을 통한 정규화가 필요할 때가 있다.

이상값(Outlier)

□ 이상값 검출 이유

- 실험이나 측정에서 발생하는 오차를 없앨 수는 없기 때문에, 실험이나 측정의 횟수를 늘려서 오차의 평균값이나 분산을 줄여나가는 방법이 있다. 그러나 이 방법은 시간과 비용이 증가하게 되므로 좋은 방법은 아님
- 이상값을 검출하여 분석한 후 제거할 수 있는 근거를 찾는 것이 필요

□ 이상값이 생기는 원인

- 측정하거나 관찰할 때 잘못 기록
- 컴퓨터에 잘못 입력하여 발생하는 실수(mistake)
- 다른 자료에 있는 관찰치가 섞여서 발생하는 자료의 오염(contamination)
- 원래 가지고 있는 자료의 고유 변동성(inherent variability)

이상값 검출 방법 (Outlier detection method)

□ R 예제코드

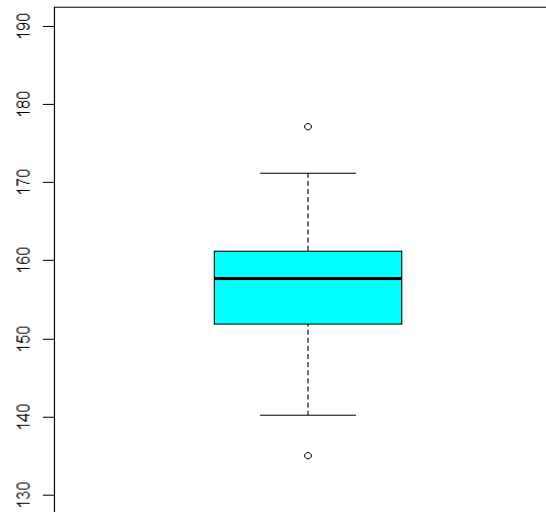
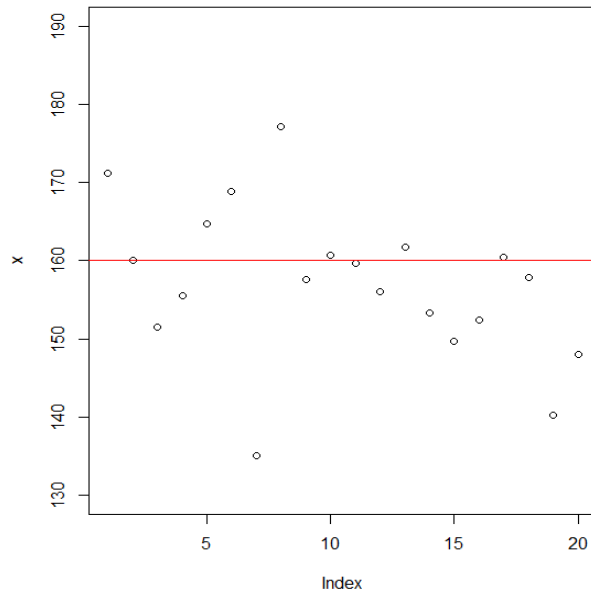
- 난수 생성 : 평균 160, 표준편차 9 인 정규분포에서 20개 난수 생성
- 이상값의 위치와 값을 확인하시오.

```
set.seed(24579)
# 난수 생성
x<-rnorm(20, 160, 9)
par(mfrow=c(1,2))
plot(x, ylim=c(130, 190))
abline(160, 0, col= ' red ' )
boxplot(x, ylim=c(130, 190), col=5)

# 이상값의 위치와 값을 확인
boxplot.stats(x)$out
oi<-which(x %in% boxplot.stats(x)$out)
x[oi]
```

이상값 검출 방법 (Outlier detection method)

□ R 결과



```
> boxplot.stats(x)$out  
[1] 135.0184 177.2016  
> ol<-which(x %in% boxplot.stats(x)$out)  
> x[ol]  
[1] 135.0184 177.2016
```

이상값 처리 (Outlier treatment)

□ 제거 (elimination)

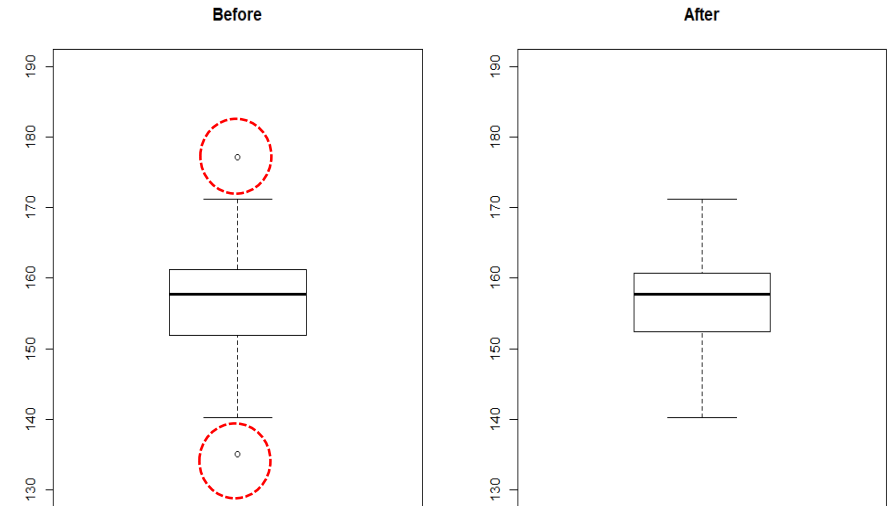
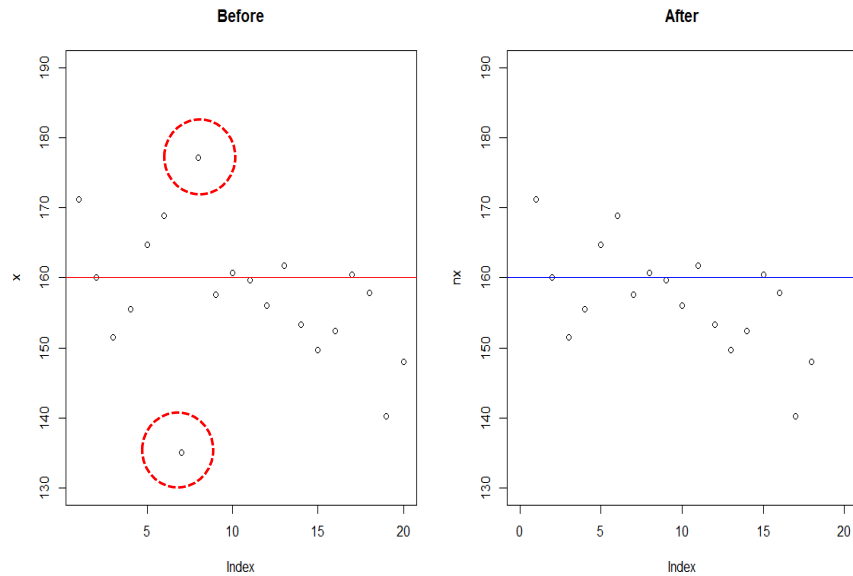
- 이상값을 제거한 후 양질의 자료로 분석과 해석을 하는 것이 좋은 결과를 도출
- 그러나 가능하면 원인을 찾아내고 제거하도록 하여, 잘못을 다시 반복하지 않게 하는 것이 좋다.

```
# 자료에서 이상값을 제거
nx<-x[-oi]

par(mfrow=c(1,2))
plot(x, ylim=c(130, 190), main="Before")
abline(160, 0, col="red")
plot(nx, xlim=c(0, 20), ylim=c(130, 190), main="After")
abline(160, 0, col="blue")
boxplot(x, ylim=c(130, 190), main="Before")
boxplot(nx, ylim=c(130, 190), main="After")
```

이상값 처리 (Outlier treatment)

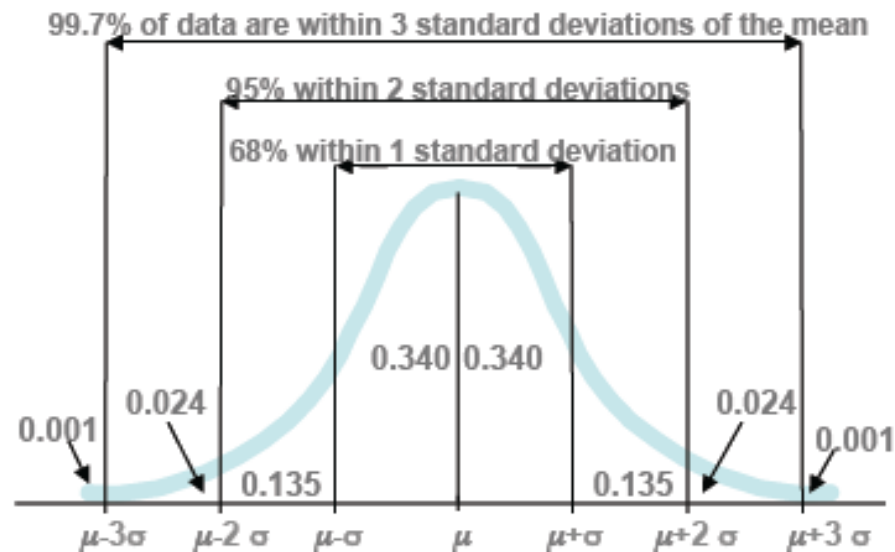
□ R 결과



이상값 판정 방법

□ 3-Sigma

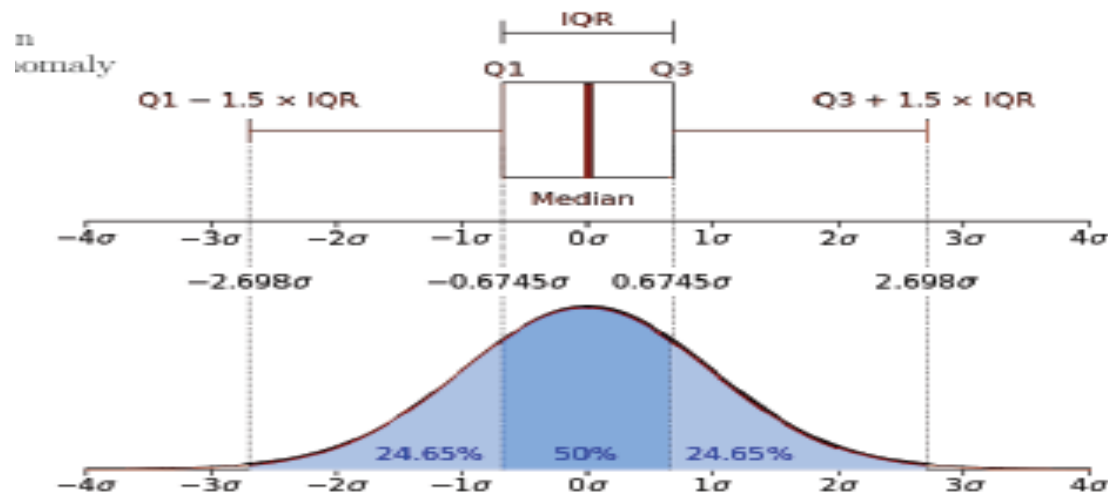
- 일변량 자료들 중 $\mu \pm 3\sigma$ 를 벗어나는 것들을 비정상이라 규정한다. (정규분포를 따르는 자료의 99.7%가 속한다는 사실에 기반한다.)
- 문제점 : 데이터에 이상값이 포함되면 평균은 정상자료 평균과 크게 다르며 sigma를 잘 추정하기 어렵다. 그렇기 때문에 평균 대신 중앙값으로 대체할 때도 있다.



이상값 판정 방법

□ Box-plot 방법

- Q3 : 제 3사분위수 (75% percentile)
- Q1 : 제 1사분위수 (25% percentile)
- IQR : Q3-Q1 (Inter Quartile Range)
- $X > Q3 + 1.5 \times \text{IQR}$ or $X < Q1 - 1.5 \times \text{IQR}$ 이면 x 를 이상값으로 규정한다.

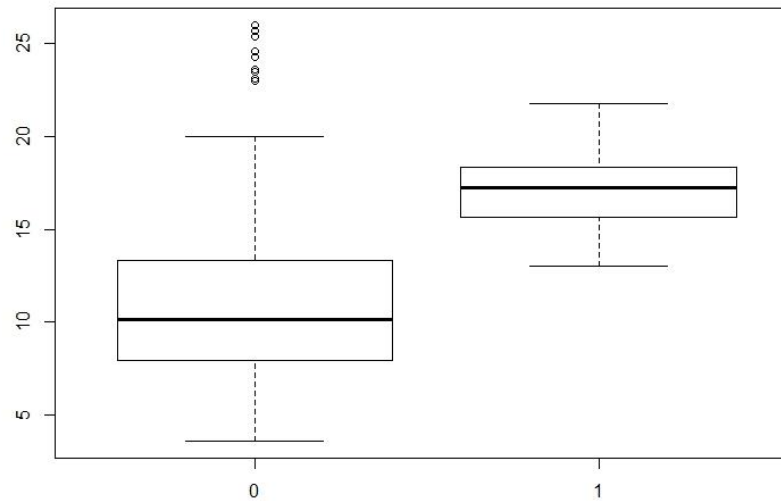


이상값 처리 방법론

□ 이상값 처리 방법론

- 제거
- 자료의 하한 또는 상한 값으로 대체
- 그러나 이상값은 함부로 제거하면 안된다.

(예 : 온도에 대한 양품(0)과 불량품(1)의 상자그림)



결측값의 처리 및 대체

□ 결측값의 종류

- 결측값 발생에 대한 구조적 형태를 아는 것은 분석과 결과 해석에 중요한 영향을 준다.
- Little and Rubin(2002), 결측값이 발생하는 구조적 형태에 따라 3가지로 분류

□ 완전 임의 결측(missing completely at random MCAR)

- 자료 입력시 실수로 빠뜨리고 입력하는 경우

□ 임의 결측(missing at random MAR)

- 결측값이 분석하려는 변수 혹은 정보와 관련이 있는 경우
- 남성이 여성보다 체중이나 나이에 더 잘 응답하는 경향이 있으며, 이때 여성의 자료에서 결측값이 발생

□ 비임의 결측(missing not at random MNAR)

- 결측값이 분석하고자 하는 정보나 변수와 관계가 없는 경우

결측값을 처리하는 방법

□ 삭제하는 방법

- 결측값이 10% 이하일 경우 : 데이터를 지우거나 대체를 한다.
- 결측값이 10% 이상일 경우 : 변수를 제거하거나 대체한다.
- 대체법에는 평균 대체법, 단순확률 대체법, KNN 대체법, 다중 대체법이 있다.

□ 대체하는 방법

- 동일값 비율이 90% 이상일 경우, 변수 제거
- Near Zero Variance (0에 가까운 분산)
- 예를 들어, 데이터 1000개가 있는데 이 중 990개에서 변수 A의 값이 1이라고 하자, 그러면 변수 A는 서로 다른 관찰을 구분하는데 별 소용이 없다. 따라서 데이터 모델링에서도 그리 유용하지 않다. 이런 변수는 분산이 0에 가깝기 때문에 변수를 제거한다.

결측값 처리 방법론

□ 결측값 처리 방법론(제거)

- 이상값과 마찬가지로 자료가 많지 않은 경우는 권장하지 않으며, 단순 제거는 바람직하지 않다.

	1	2	3	4	5	6
1						NA
2	NA					
3						
4		NA				
5				NA		
6						NA
7						

- 너무 많은 데이터(정보)의 손실이 생긴다.

결측값 처리 방법론

□ 결측값 처리 방법론(대치법 : 평균 대치법)

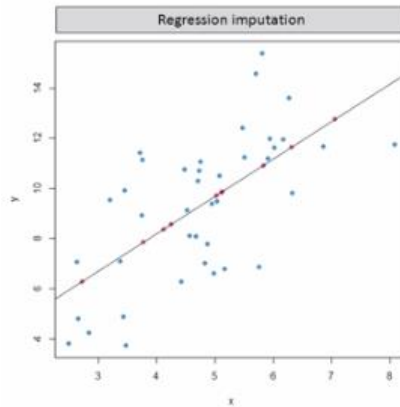
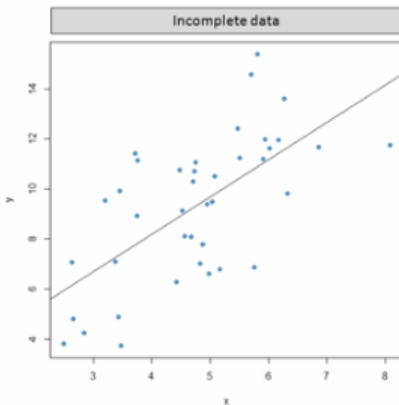
- 관측 또는 실험되어 얻어진 자료의 적절한 평균값으로 결측값을 대치하는 방법이다.
- 평균 대치법은 사용하기가 간단하고 제거하는 방법에 비해 효율성이 향상된다.
- 그러나 관측된 자료를 토대로 한 추정값으로 결측값을 대치함으로써 통계의 표준오차가 과소 추정되는 문제가 있다.

X1	X2	X3	X4
15	234	10	42
23	232	5	53
46	152	NA (7로 대치)	46
23	234	6	32
16	345	7	54

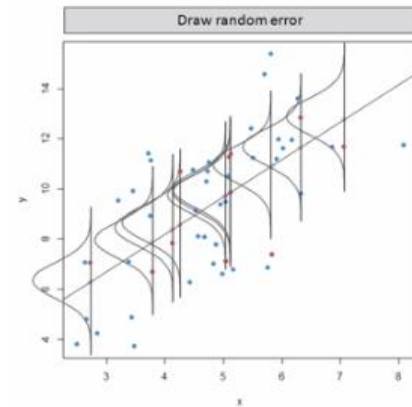
결측값 처리 방법론

□ 결측값 처리 방법론(대치법 : 단순 확률 대치)

- 평균 대치법에서 추정량 표준오차의 과소 추정문제를 보완하고자 고안한 방법이다.
- 평균 대치법에서 관측된 자료를 토대로 추정된 통계량으로 결측값을 대치할 때 어떤 적절한 확률 값을 부여한 후 대치하는 방법이다.
- 추정량의 표준오차가 과소 추정되는 문제는 보완 되지만 간단 문제를 제외한 대부분의 경우에 추정량의 표준오차 계산자체가 어려운 문제가 있다.



회귀분석으로 구한 결측
값 예측 (붉은 점)

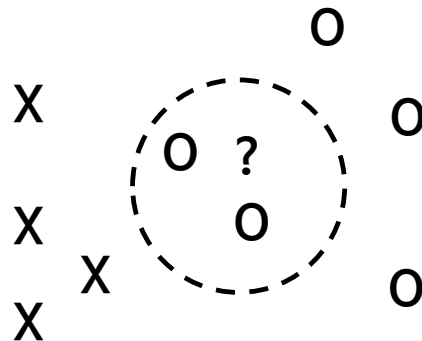


각 예측값에 대한 Random
error을 적용

결측값 처리 방법론

□ 결측값 처리 방법론(대치법 : KNN 대치)

- 예측하고자하는 데이터로부터 가장 가까운 k개 이웃을 찾은 뒤 이들 이웃으로부터 예측하고자 하는 데이터의 분류를 정하는 방법이다.
- 예를 들어, X,O는 각 점의 분류를 의미하고 ?로 표시된 점은 분류를 예측해야 하는 점이다.
K=2인 경우 ?로부터 가장 가까운 점은 점선으로 표시한 원 안에 있는 O 점 두 개다. 따라서 ?의 분류는 KNN알고리즘에서 O로 예측한다.



결측값 처리 방법론

□ 결측값 처리 방법론(대치법 : 다중 대치)

- 추정량 표준오차의 과소 추정 또는 계산의 난해성의 문제를 보완 할 수 있는 방법이다.
- 다중 대치법은 단순 대치법(평균 대치, 단순 확률 대치)을 한번 하지 않고 m번의 대치를 통한 m개의 가상적 완전한 자료를 만들어서 분석하는 방법으로 다음과 같이 3가지 단계로 구성되어 있다.

1) 대치(Imputations step)

2) 분석(Analysis step)

3) 결합(Combination step)