1. 기초통계량

A. 데이터 분석 과정

1) 정의 단계 : 문제의 정의

고객으로부터 최대한의 정보를 얻어내야 함

2) 분해 단계 : 작은 단위로 분할한 후에 단계별로 해결

확보한 데이터를 분할된 단위에 맞추어 수집하거나 재구성

고객이 제공한 문제의 본질을 이해하고 분석 가능한 작은 단위로 문제를 분할한 후에 분석 수행

문제의 분해는 결과에 대한 예측을 기반으로 실행함

3) 평가 단계

주어진 문제와 고객이 알고자 하는 것을 기준으로 현재의 시점에서 결과를 평가하는 단계

4) 결정 단계

평가가 완료된 후 분석가의 결정을 전달하는 과정

데이터 분석 모델을 확정하고 데이터를 분석하여 최종적인 분석가의 의견을 확정하는 단계

5) 반복 단계

새로운 자료나 상황이 발생할 경우 이미 실행한 단계를 다시 수행해야 함

B. 데이터의 유형

1) 범주형 데이터(Categorical Data)

사전에 정해진 특정 유형으로 분류되는 데이터

ex) 방의 크기 : 대,중,소

① 명목형 : 값들 간의 크기 비교가 불가능

ex) 정치적 성향(좌파, 우파), 성별 등

② 순서형 : 대, 중, 소와 같이 값에 순서를 매길 수 있는 경우

ex) 성적 데이터, 방의 크기

2) 연속형 데이터(Continuous Data) : 정량적 데이터

① 등간척도 : 섭씨온도, 화씨온도, 시간 등

② 비율척도 : 키, 몸무게, 점수, 관찰빈도 등

변수 유형	자료 유형	예
질적변수	명목형 변수	성별, 혈액형
	순서형 변수	학력, 설문문항
양적변수	이산형 변수	가족수, 수강과목 수
	연속형 변수	키, 몸무게

C. 데이터 시각화 방법

1) 기초통계량(평균, 분산 등) : 전체 데이터의 특성을 표현하는 수치

2) 산포도(산점도):전체 데이터가 어떤 특징을 가지는지 보여줌
   x축 독립변수(영향을 미치는 변수, 원인)
   y축 종속변수(결과)

3) 히스토그램 : 데이터를 구간별로 나누어 도수를 표시함으로써 특징을 보여줌

D. 기초통계량

1) 최대값, 최소값

2) 최빈값 : 가장 많이 관찰된 값

3) 평균값

4) 중앙값

5) 표준편차

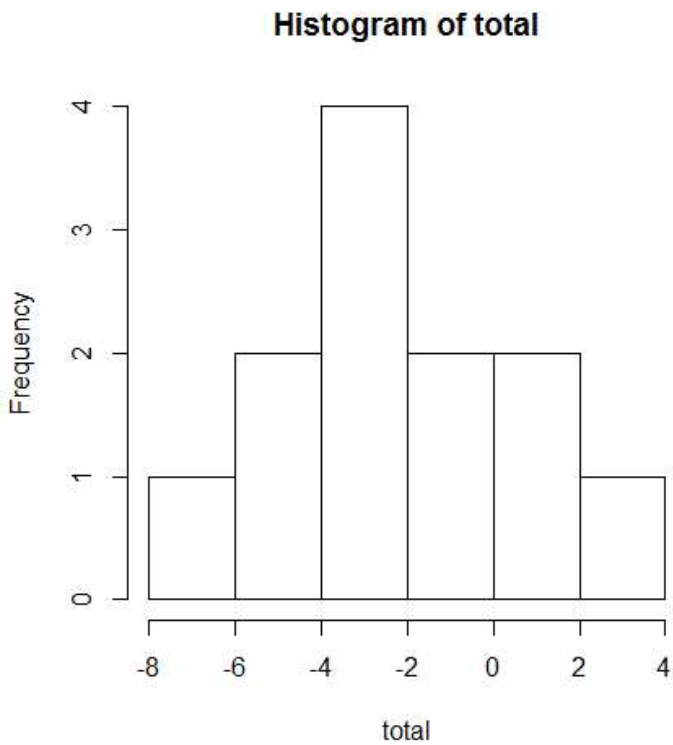
6) 사분위수

	R함수	설명
자료의 갯수	length()	
최소값	min()	
최대값	max()	
범위	range()	
최빈값		자료 중 빈도수가 가장 많은 값
평균	mean()	
중앙값	median()	자료를 순서대로 나열했을 경우 중앙에 있는 값
표준편차	sd()	평균을 중심으로 자료가 퍼진 정도
제1사분위수	quantile()	자료를 순서대로 나열했을 때 25% 위치의 값
제3사분위수		자료를 순서대로 나열했을 때 75% 위치의 값
사분위수 범위	IQR()	제3사분위수-제1사분위수

E. 실습예제

1. 데이터를 R로 읽어들인다.
2. 읽어들인 데이터로 그래프를 그린다.
3. 읽어들인 데이터의 기초통계량을 구한다.

```
#데이터를 R로 로딩
flour <- c( 3, -2, -1,  0,  1, -2)  # 밀가루 사용
diet  <- c(-4,  1, -3, -5, -2, -8)  # 다이어트약 사용
total <- c(flour, diet)             # 12명의 데이터
#히스토그램
hist(total)
```



```
#ann=F 축의 라벨을 표시하지 않음
#density() 확률밀도 그래프
plot(density(flour),  xlim=c(-8,8),  ylim=c(0,0.2),  lty=1,
ann=F)
par(new=T) #2개의 그래프를 하나에 출력
plot(density(diet),  xlim=c(-8,8), ylim=c(0,0.2), lty=2)
legend(4, 0.2, c("flour","diet"), lty=1:2, ncol=1)   # 범례
```

```
# 밀도 그래프(density)
# 히스토그램은 막대 구간의 너비에 따라 모양이 달라질 수 있음
# 밀도 그래프는 막대의 너비를 가정하지 않고 모든 점에서 데이터의
밀도를 추정하는
# 커널 밀도 추정kernel density estimation 방식을 사용하여 이러
한 문제를 보완함
# density(커널밀도를 추정할 데이터)
# rug(숫자벡터) : 그래프의 x축에 데이터를 1차원으로 표시
```

```
plot(density(iris$Sepal.Width))
```

```
#히스토그램 위에 밀도 그래프를 선으로 표시할 수 있음
hist(iris$Sepal.Width, freq=F)
lines(density(iris$Sepal.Width))
```

```
#밀도 그래프에 rug() 함수를 이용하여 실제 데이터의 위치를 x축 위
에 표시한 그래프
#jitter() 데이터의 중첩
plot(density(iris$Sepal.Width))
rug(jitter(iris$Sepal.Width))
```

```
#Jitter : 같은 값을 가지는 데이터가 같은 좌표에 겹쳐서 표시되지 않도록
#  데이터에 약간의 노이즈를 추가하는 방법
```

```
iris$Sepal.Width
```

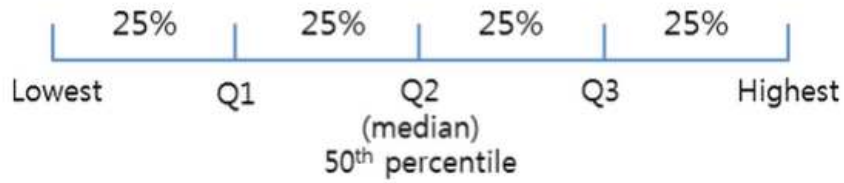
```
jitter(iris$Sepal.Width)
```

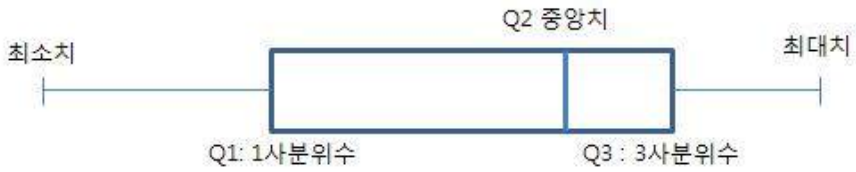
상자수염그림 – 두 그룹의 분포를 비교할 목적

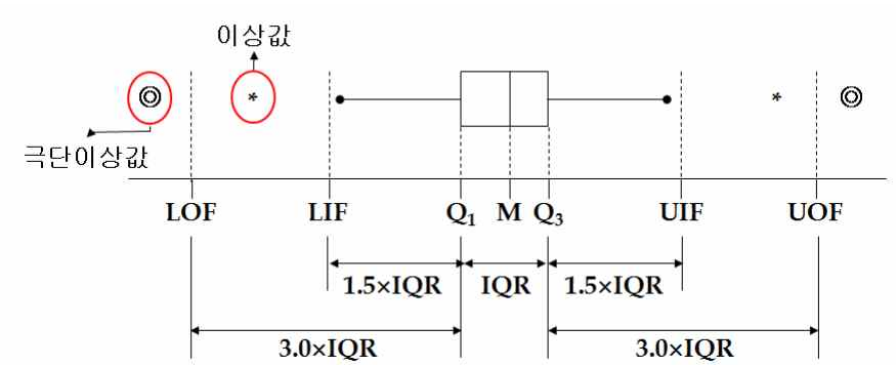
사분위수 : 자료를 순서대로 정렬한 후 4등분 한 것

0%(Lowest), 25%(Q1), 50%(Q2), 75%(Q3), 100%(Highest)

Q3-Q1 : 사분위수 범위(Interquartile Range : IQR) 상자의 길이







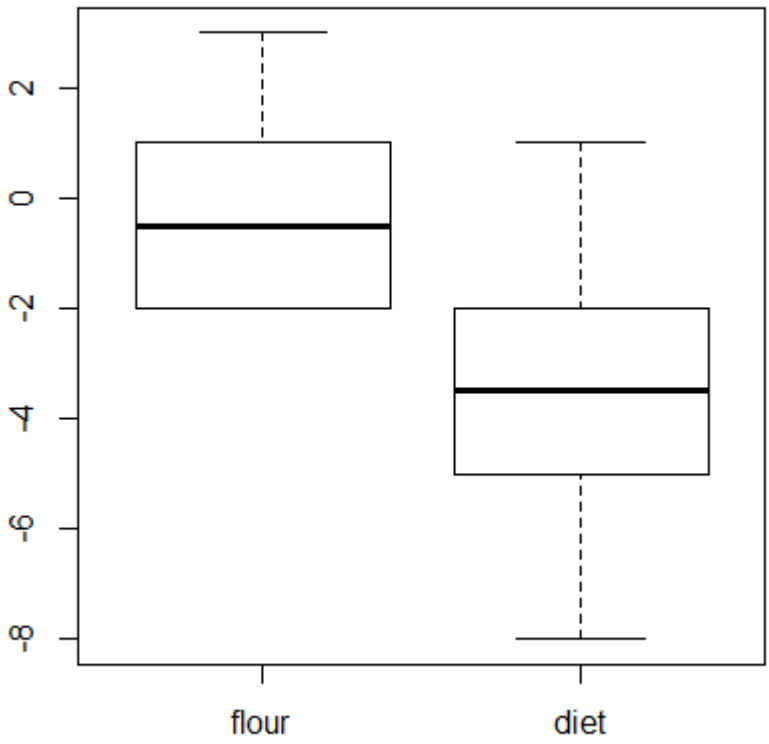
LIF(Lower Inner Fence : LIF) 하 내부울타리 Q1 - 1.5×IQR

UIF(Upper Inner Fence : UIF) 상 내부울타리 Q3 + 1.5× IQR

LOF(Lower Outer Fence) 하 외부울타리 Q1 – 3.0 x IQR

UOF(Upper Outer Fence) 상 외부울타리 Q3 + 3.0 x IQR

```
boxplot(flour, diet, names=c("flour", "diet"))
```



기초 통계량 계산

```
#합계
sum(total)
#quantile 분위수
quantile(total)
#   0%   25%   50%   75%  100%
# -8.00  -3.25   -2.00   0.25  3.00
fivenum(total)                  # min, 1Q, median, 3Q, max
#fivenum과 quantile 계산 방식이 다름
# -8.0 -3.5 -2.0  0.5  3.0
cor(flour, diet)                # 상관계수


summary(total)
```

커피 판매량 데이터

```
cafe <- read.csv("d:/data/cafe/data.csv")
# 자료 정렬
sort(cafe$Coffees)
```

```
3 4 4 4 4 5 6 8 8 10 11 11 13 14 16 16 18 20 20 21 21 22 23
23 24 24 25 25 26 27 27 27 28 29 30 30 31 31 31 32 33 33 34
35 35 41 48
```

1) 최대, 최소값

```
#정렬된 값 중 첫번째 값
sort(cafe$Coffees)[1]

#내림차순 정렬
sort(cafe$Coffees, decreasing=TRUE)
#내림차순 정렬된 값 중 첫번째 값
sort(cafe$Coffees, decreasing=TRUE)[1]
#최소값
min( cafe$Coffees )
#최대값
max( cafe$Coffees )
#하루 주문량은 3~48잔임을 알 수 있음
```

2) 최빈값(mode) : 가장 많이 관찰된 값, 가장 빈도가 높은 값

```
#seq(0,50,by=10) 0~50까지 10씩 증가
#right=F : 마지막 값은 선택하지 않음
table(cut(cafe$Coffees, breaks=seq(0, 50, by=10), right=F))
#0~9잔 : 9, 10~19잔 8, 20~29잔 17
```

```
[0,10) [10,20) [20,30) [30,40) [40,50)
   9      8     17     11      2
```

```
ca <- cafe$Coffees
stem(ca)
#줄기-잎 그림 : 자료를 순서대로 나열한 후 적당한 단위로 나눠 줄기 부분을 만들고 각 값을 줄기 부분에 붙인 그림
```

```
  The decimal point is 1 digit(s) to the right of the |
```

십의 자리가 줄기, 일의 자리가 잎

```
  0 | 34444
  0 | 5688
  1 | 01134 => 10, 11, 11, 13, 14
  1 | 668
  2 | 001123344
  2 | 55677789
  3 | 001112334
  3 | 55
  4 | 1
  4 | 8
```

```
ca
#각 잔별로 빈도수가 출력됨 - 3잔은 1회, 4잔은 4회 ...
table(ca)
max( table(ca) )

# 최빈값은 4임을 알 수 있다.
```

```
41 33 34 27 20 23 32 31 30 27 30 27 26 24 18 22 21 28 23 31 29
48 25 31 25 35 33 35 16 24 20 11 21 8 8 4 4 3 5 6 4 13 4 16 14
10 11


ca
 3  4  5  6  8 10 11 13 14 16 18 20 21 22 23 24 25 26 27 28 29
30 31 32 33 34 => 커피 판매량
 1  4  1  1  2  1  2  1  1  2  1  2  2  1  2  2  2  1  3  1  1  2
3  1  2  1
35 41 48
 2  1  1  => 빈도수

4 => 최빈수는 4이다.
```

3) 평균값

```
ca <- cafe$Coffees
#평균값 계산
mean( ca )
```

```
21.5106382978723
```

```
# ca 변수에 결측값을 덧붙임 (NA=Not Available)
ca <- c(ca, NA)
tail(ca, n=5)
```

```
16 14 10 11 <NA>
```

```
#결측값이 있으므로 평균값이 계산되지 않음
mean( ca )

#결측값을 제외하고 평균 계산
mean( ca, na.rm=T )
```

```
<NA>
21.5106382978723
```

4) 중앙값, 중위수(median) - 자료 전체의 중심 위치값

```
ca <- cafe$Coffees
ca

#오름차순 정렬
sort(ca)

#중앙값
median( ca )
```

```
41 33 34 27 20 23 32 31 30 27 30 27 26 24 18 22 21 28 23 31
29 48 25 31 25 35 33 35 16 24 20 11 21 8 8 4 4 3 5 6 4 13 4
16 14 10 11

3 4 4 4 4 5 6 8 8 10 11 11 13 14 16 16 18 20 20 21 21 22 23
23 24 24 25 25 26 27 27 27 28 29 30 30 31 31 31 32 33 33 34
35 35 41 48

23
```

5) 분산, 표준편차

: 각 자료들이 평균에 대해서 평균적으로 얼마나 떨어져 있는가?
분산(variance) : 자료가 흩어져 있는 정도

관찰값들이 평균값으로부터 얼마나 많이 퍼져 있는지를 파악하기 위한 것이 분산

분산을 알기 위해서는 먼저 평균을 알아야 하고, 각각의 관찰값들과 평균 사이의 거리(distance)를 재기 위해 관찰값에서 평균을 빼게 된다. 그런데 어떤 관찰값들은 반드시 평균 이하에 존재하고 있고, 이들의 존재로 인해 양수 값들과 음수 값들이 혼재하게 된다. 그리고 이들을 모두 합칠 경우 결과는 0이 나온다.

ex) 1 2 3 4 5 6 7 평균 4

(1-4)+(2-4)+(3-4)+(4-4)+(5-4)+(6-4)+(7-4)

=-3 -2 -1 + 0 + 1 + 2 + 3

=0

이 문제를 해결하기 위해 각각의 편차를 제곱한 값((편차제곱, squares of deviation))을 모두 더한 후 전체 관찰값의 개수만큼 나눈다. 즉 편차제곱을 가지고 평균을 구하는 것이다. 이것이 분산이다.

통계처리를 할 때 모집단 전체에 대한 전수조사보다는 표본을 이용하는 경우가 대부분임.
모집단의 분산을 구할 때는 n으로 나누고
표본집단의 분산을 구할 때는 n으로 나누지 않고 n-1로 나눈다.(표본분산)

ex) 1 2 3 4 5 6 7 평균 4

$(1-4)^2+(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2+(7-4)^2$

=> 9+4+1+0+1+4+9 / 6

=> 28 / 6

=> 4.666667

```
height <- c(164, 166, 168, 170, 172, 174, 176)

#평균값
mean(height)
```

```
170
```

```
#중앙값
median(height)
```

```
170
```

```
#편차(deviation) : 개별 관찰값과 평균과의 차이
#편차값을 모두 더하면 0이 되므로 의미가 없고 제곱을 해야 한다.
height.dev <- height - mean(height)
height.dev
sum(height.dev)
```

```
-6 -4 -2 0 2 4 6
```

```
#분산(variance, 편차 제곱의 평균)
var( height )
```

```
18.6666666666667
```

키의 평균은 170cm이고 각 자료들은 평균으로부터 평균적으로 $19cm^2$ 정도 떨어져 있음(분산). 평균의 단위는 길이이고 분산의 단위는 면적이므로 단위가 맞지 않음. 따라서 단위를 맞추기 위해 제곱값인 분산의 제곱근을 구한 값을 표준편차(Standard Deviation)라고 함

```
#표준편차
sd( height )
```

4.32049379893857

모집단의 분산, 표준편차와 달리 R에서 계산하는 분산,표준편차는 분모가 N이 아닌 N-1로 계산됨. 모집단인 경우에는 분모가 N이고 표본집단의 경우는 N-1로 계산. 통계분석은 대부분 표본집단을 대상으로 처리되므로 N-1로 계산함

변동계수 구하기(커피와 주스 중 어떤 음료의 판매량 변동폭이 더 큰가?)

표준편차를 산술평균으로 나눈 값(측정 단위가 서로 다른 자료를 비교할 경우 사용)

```
coffee <- cafe$Coffees
juice <- cafe$Juices

#커피 판매량 평균값
( coffee.m <- mean( coffee ) )
```

21.5106382978723

```
#커피 판매량의 표준편차
( coffee.sd <- sd( coffee ) )
```

11.080480958847

```
#쥬스 판매량 평균값
( juice.m <- mean( juice ) )
```

4.93617021276596

```
#쥬스 판매량의 표준편차
( juice.sd <- sd( juice ) )
```

```
3.70313765503484
```

```
# 커피 판매량의 변동계수
( coffee.cv <- round( coffee.sd / coffee.m, 3) )
```

```
0.515
```

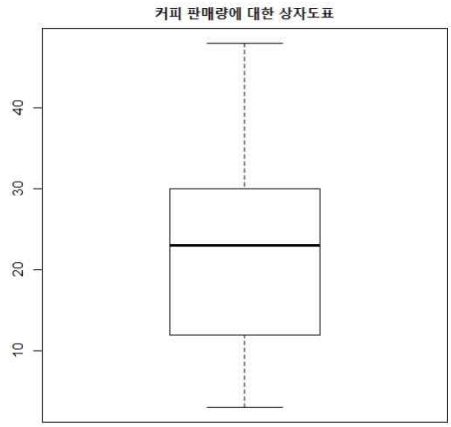
```
# 쥬스 판매량의 변동계수(표준편차를 평균으로 나눈 값)
( juice.cv <- round( juice.sd / juice.m, 3) )

#표준편차값을 보면 커피가 훨씬 크지만 변동계수를 볼 때 쥬스가 더 크다. 즉, 쥬스 판매량의 변동폭이 더 크다는 것을 알 수 있다.
```

```
0.75
```

6) 사분위수 범위와 상자도표

```
quantile(coffee)
#25%가 되는 값(제1사분위수, Q1) 12, 50%가 되는 값(제2사분위수, Q2) 23, 75%가 되는 값(제3사분위수, Q3) 30, 100%가 되는 값(제4사분위수, Q4)

#사분위수 범위(InterQuartile Range, Q3 - Q1)
IQR(coffee)

boxplot(coffee, main="커피 판매량에 대한 상자도표")
#상자의 아랫변 Q1, 상자중앙의 굵은 선 Q2, 상자의 윗변 Q3

d<-matrix(c(coffee,juice),47,2)
d
win.graph()
boxplot(d,names=c('coffee','juice'))

boxplot(coffee, juice, names=c('커피판매량','주스판매량'))
```

```
  0%   3
 25%  12
 50%  23
 75%  30
100%  48

18
```


커피 판매량에 대한 상자도표
10
20
30
40

2. 가설 검정

A. 통계 분석

1) 모집단과 표본

① 모집단 : 우리가 알고자 하는 대상 전체, 조사 대상의 범위

② 표본 : 모집단으로부터 조사하기 위해 선택된 조사대상

2) 전수조사와 표본조사

① 전수조사: 모집단을 구성하는 대상 전부를 조사하는 것

가장 정확하지만 비용과 시간이 많이 들게 됨

전수조사가 불가능한 경우도 있음(예를 들어 감기약의 경우 모두 복용을 해야만 효과를 알 수 있음)

② 표본조사 : 표본을 대상으로 조사

3) 통계 분석 기법

① 어떤 그룹,집단,형태 등의 차이를 검정

1개, 2개 또는 그 이상의 데이터 차이가 있다고 볼 수 있는지를 검정하는 것

독립표본 t검정, 대응표본 t검정, ANOVA 등

② 요소와 요소간의 인과관계(상관관계)를 파악

상관분석 – 변수와 변수 사이의 직선 관계를 상관계수를 이용해서 분석

회귀분석 – 종속변수와 독립변수간의 관계를 모형화하여 분석

B. 가설 검정

1) 가설 검정의 기초

① 과학 분야에서의 증명 : 반증법에 의거해 증명

"모든 사람은 정직하다"라는 명제가 있을 때 이 명제가 참인지 거짓인지를 확인하는 접근법에는 2가지가 존재

(1) 모든 사람을 일일이 조사해서 정직한지 확인하는 방법

(2) 정직하지 못한 사람(사례)을 하나 찾아내 명제가 거짓임을 입증하는 방법

위 2가지 방법 중 어느 것이 효율적일까? (현실적으로 가능한 방법일까?)

1000명이 정직함을 확인했을 때, 1001명째 사람이 정직하다고 확언할 수 있을까?

② 귀무가설(H0)과 대립가설(Ha)

우리가 알고 싶은 명제(주로 A와 B는 다르다, A와 B는 차이가 있다 등)는 대립가설로 두고, 그 반대인 귀무가설(주로 A와 B는 같다, A와 B는 차이가 없다 등)이 기각됨을 보임으로서 내가 입증해 보이고 싶은 명제를 증명

2) 독립표본 t-검정

서로 독립된 두 집단간의 평균의 차이가 통계적으로 유의미한지 비교하고자 할 때 사용

즉, 서로 독립된 두 집단에 대해 각 집단별 특정 연속형 변수 평균값이 서로 차이가 있는지 없는지를 통계적으로 검정할 때 사용되는 기법

예) 전체 응답자 중 남자와 여자 사이의 연령은 차이가 있는가?

3) 대응표본 t-검정(Paired-samples t-test)

서로 동일한 모집단에서 추출된 두 표본에 대해 특정 연속형 변수 평균값이 서로 차이가 있는지, 없는지를 통계적으로 검정할 때 사용되는 기법

예) 한 회사에서 자사가 개발한 한 달간의 식이요법 프로그램이 효과가 있는지 여부를 분석하고자 함

4) 일원배치 분산분석(One-way ANOVA)

세 개 이상의 집단간의 평균의 차이가 통계적으로 유의미한지 비교하고자 할 때 사용

예) 학력수준에 따라 직무만족도의 수준은 차이가 있는가?

5) 가설검정의 정리

분석의 대상이 되는 두 변수의 특징에 따라, 다음과 같이 다른 검정기법을 적용해야 함

기법	대상 변수A	대상 변수B	적용 예
카이제곱 검정	이산형	이산형	성별과 구매여부 사이에 유의한 관계가 있는가? 성별과 결혼유무 사이에 유의한 관계가 있는가?
독립표본 t검정	이산형 (2그룹)	연속형	체중과 구매여부 사이에 유의한 관계가 있는가? (구매자와 비구매자의 평균 체중이 크게 다른가?) 성별에 따른 평균 취업률의 차이가 있는가?
대응표본 t검정	이산형 (2그룹/ Pair)	연속형	보충수업 후 성적의 향상이 있는가?
일원배치 분산분석	이산형 (3그룹 이상)	연속형	체중과 고객등급 사이에 유의한 관계가 있는가? (고객등급에 따라 평균 체중이 크게 다른가?) 거주지역에 따른 평균소득액의 차이가 있는가?

C. 실습예제

1) 카이제곱 검정(Chi-Square Test)

두 범주형 변수(범주로만 분류되고 수치적으로 측정되지 않는 자료)가 서로 상관이 있는지 판단하는 통계적 검정 방법

ex) 학력, 성별, 직업의 만족도 등

귀무가설 : child1과 child2 두 데이터는 차이가 없다.

대립가설 : child1과 child2 두 데이터는 차이가 있다.

```
child1<-c(5,11,1)
child2<-c(4,7,3)
Toy<-cbind(child1,child2)
rownames(Toy)<-c("car","truck","doll")
Toy
```

```
chisq.test(Toy) #카이제곱 검정
```

```
        Pearson's Chi-squared test
data:  Toy
X-squared = 1.7258, df = 2, p-value = 0.4219
```

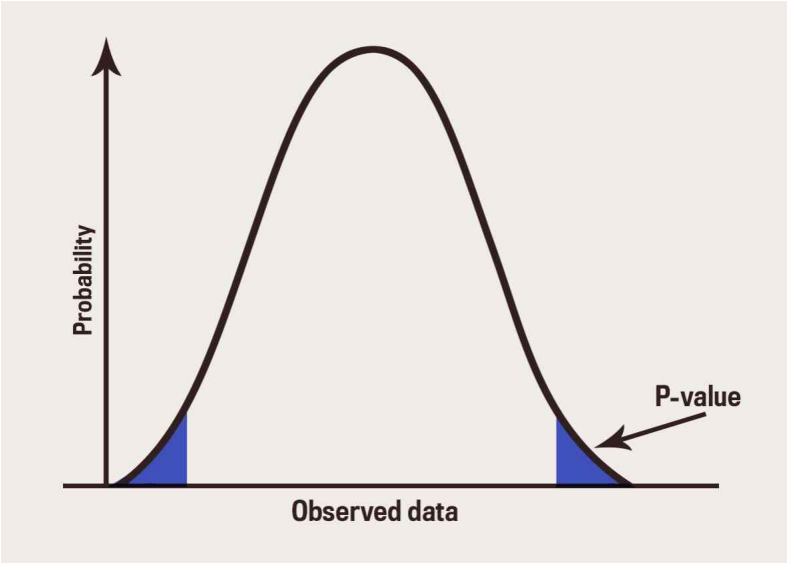
p-value가 0.05보다 크므로 귀무가설을 기각할 수 없다. 따라서 child1과 child2 두 데이터는 통계적으로 차이가 없다.

```
Warning message:
In chisq.test(Toy) : 카이제곱 approximation(근사치)은 정확하지 않을수도 있습니다
```

=> 빈도가 5보다 작은 셀이 전체의 20% 이상인 경우의 경고 메시지, 이런 경우에는 피셔 검정(Fisher's exact test)을 실시해야 함

-귀무가설(null hypothesis, 영가설, H_0) : 기존의 가설
-대립가설(H_1) : 내가 주장하고 싶은 새로운 가설
귀무가설과 대립가설은 정반대로 설정되어야 한다.
ex) 귀무가설 : 담배는 수명에 영향을 주지 않는다.
대립가설 : 담배는 수명을 단축시킨다.
ex) 강아지의 평균수명은 13년이라고 하는데 실제로 그런지 검정한다면
귀무가설 : 강아지의 수명은 13년이다.
대립가설 : 강아지의 수명은 13년이 아니다.
ex) 어떤 제품의 불량률이 3% 이하라고 한다. 그런데 불량에 대한 항의전화가 많이 와서 실제 불량률이 3%인지 검정한다면
귀무가설 : 제품의 불량률은 3% 이하이다.
대립가설 : 제품의 불량률은 3%보다 크다.

p-value : 0에 가까울수록 좋다. 미리 정해진 유의수준(일반적으로 0.05)보다 작으면 귀무가설을 기각하고 대립가설을 채택할 수 있다.



* 피셔 검정

표본수가 적거나 데이터의 분포가 치우친 경우에 적용됨

```
fisher.test(Toy)
```

```
  Fisher's Exact Test for Count Data

data:  Toy
p-value = 0.5165
alternative hypothesis: two.sided
```

귀무가설 : child1과 child2 두 데이터는 차이가 없다.

대립가설 : child1과 child2 두 데이터는 차이가 있다.

결론 : p-value가 0.05보다 크기 때문에 child1과 child2 두 데이터는 차이가 없다.(대립가설을 기각한다.)

2) t검정

표본으로부터 추정된 분산이나 표준편차를 가지고 가설을 검정하는 방법

귀무가설 : 두 모집단은 평균간의 차이가 없다.
대립가설 : 두 모집단은 평균간의 차이가 있다.

① 표본이 1개인 경우

전체 모집단에 대한 정보가 없을 때 표본이 과연 모집단으로부터 나온 것인지를 판단할 때 사용

A회사의 건전지 수명시간이 1000시간일 때 무작위로 뽑은 10개의 건전지에 대해 수명은 다음과 같다.
980,1008,968,1032,1012,1002,996,1017 샘플이 모집단과 다르다고 할 수 있는가?

귀무가설 : 건전지의 수명은 1000시간이다.
대립가설 : 건전지의 수명은 1000시간이 아니다.

먼저 데이터분포가 정규분포인지 아닌지를 확인하기 위해 Shapiro-Wilk 검정을 실시한다.

Shapiro-Wilk 검정의 가설
귀무가설 : 자료가 정규분포를 따른다.
대립가설 : 자료가 정규분포를 따르지 않는다.

```
a<-c(980,1008,968,1032,1012,1002,996,1017)

shapiro.test(a)
```

```
        Shapiro-Wilk normality test

data:  a
W = 0.97706, p-value = 0.9469
```

p-value가 0.05보다 크므로 귀무가설을 기각할 수 없다. 따라서 이 자료는 정규분포를 따른다. 이 데이터에 대해 T-Test를 할 수 있다.

```
# a 샘플 데이터 벡터, mu : 비교하는 대상의 평균
# alternative : two.sided 샘플이 주어진 평균과 다르다, greater 샘플이 주어진 평균보다 크다, less 샘플이 주어진 평균보다 작다
t.test(a,mu=1000,alternative="two.sided")
```

```
        One Sample t-test

data:  a
t = 0.25891, df = 7, p-value = 0.8032
alternative hypothesis: true mean is not equal to 1000
95 percent confidence interval:
  984.7508 1018.9992
sample estimates:
mean of x
 1001.875
```

p-value가 0.05보다 크므로 귀무가설을 기각할 수 없다. 따라서 건전지의 수명은 1000시간이다.

* 어떤 학급의 수학 평균성적은 55점이었다. 0교시 수업을 시행하고나서 학생들의 시험 성적은 다음과 같았다.

58, 49, 39, 99, 32, 88, 62, 30, 55, 65, 44, 55, 57, 53, 88, 42, 39

0교시 수업을 시행한 후, 학생들의 성적은 올랐다고 할 수 있는가?

H0(귀무가설): 0교시 수업 시행 후 학생들의 성적은 오르지 않았다.
H1(대립가설): 0교시 수업 시행 후 학생들의 성적은 올랐다.

먼저 데이터분포가 정규분포인지 아닌지를 확인하기 위해 Shapiro-Wilk 검정을 실시한다.

Shapiro-Wilk 검정의 가설
귀무가설 : 자료가 정규분포를 따른다.
대립가설 : 자료가 정규분포를 따르지 않는다.

```
a <- c(58, 49, 39, 99, 32, 88, 62, 30, 55, 65, 44, 55, 57, 53, 88, 42, 39)
summary(a)
shapiro.test(a)
```

```
    Shapiro-Wilk normality test

data:  a
W = 0.91143, p-value = 0.1058
```

p-value가 0.05보다 크므로 귀무가설을 기각할 수 없다. 따라서 이 자료는 정규분포를 따른다. 이 데이터에 대해 T-Test를 할 수 있다.

```
t.test(a, mu=55, alternative="greater")
```

```
  One Sample t-test

data:  a
t = 0.24546, df = 16, p-value = 0.4046
alternative hypothesis: true mean is greater than 55
95 percent confidence interval:
 47.80855      Inf
sample estimates:
mean of x
 56.17647
```

p-value가 0.05보다 크기 때문에 0.4086이기 때문에 귀무가설을 기각할 수 없다.
0교시를 시행해도 성적은 오르지 않았다.

② 표본이 2개인 경우의 t검정

환자 10명을 상대로 혈압약을 먹었을 때와 먹지 않았을 때의 혈압을 측정하고 이 두 자료의 평균이 같다고 할 수 있는지를 검정

귀무가설 : 혈압약을 먹었을 때와 먹지 않았을 때의 혈압 차이가 없다.
대립가설 : 혈압약을 먹었을 때와 먹지 않았을 때의 혈압 차이가 있다.

```
pre<-c(13.2,8.2,10.9,14.3,10.7,6.6,9.5,10.8,8.8,13.3)
post<-c(14,8.8,11.2,14.2,11.8,6.4,9.8,11.3,9.3,13.6)
t.test(pre,post)
```

```
  Welch Two Sample t-test

data:  pre and post
t = -0.36891, df = 17.987, p-value = 0.7165
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -2.745046  1.925046
sample estimates:
mean of x mean of y
    10.63     11.04
```

p-value가 0.05보다 크므로 귀무가설을 기각할 수 없다. 따라서 혈압약을 먹었을 때와 먹지 않았을 때의 혈압 차이가 없다.

3) ANOVA (ANalysis Of VAriance, 분산분석)

비교하고자 하는 집단이 3개 이상인 경우
k개의 모집단의 평균을 이용하여 상관관계의 유무를 판단하는 기법

귀무가설 : 평균이 같다.
대립가설 : 평균이 같지 않다.

```
#샘플데이터 생성
xx<-c(1,2,3,4,5,6,7,8,9)
yy<-c(1.09,2.12,2.92,4.06,4.9,6.08,7.01,7.92,8.94)
zz<-c(1.1, 1.96, 2.98, 4.09, 4.92, 6.1, 6.88, 7.97, 9.01)
#벡터형으로 자료를 생성함
mydata<-c(xx,yy,zz)
mydata
group<-c(rep(1,9),rep(2,9),rep(3,9))
group
```

```
oneway.test(mydata~group)
```

```
  One-way analysis of means

data:  mydata and group
F = 6.2966e-06, num df = 2.000, denom df = 15.999, p-value
= 1
```

p-value가 0.05보다 크기 때문에 귀무가설을 기각할 수 없다. 즉, 평균이 같다.