1) 독립 이표본 평균 정의

독립적인 두 집단의 평균을 비교하는 것을 말한다.

독립표본 T 검정이라고도 말할 수 있다. Independent T-test (Two sample t-test)

T검정은 T-test라고도 하며 두 집단의 평균을 구하는 것이다.

두 집단을 어떻게 보느냐에 따라서 세 가지로 나눌 수 있다.

* 일표본(One Sample) T검정
* 독립표본(independent) T검정
* 대응표본(Paired) T검정

H0 영가설(귀무가설) : 두 집단 평균은 차이가 없다.

H1 대립가설(연구가설) : 두 집단 평균은 차이가 있다.

t-검정은 t값의 숫자보다 t값이 유의수준 어디쯤에 위치하는지가 더 중요하다.

t값이 유의확률 p-value에 들어가는지, 안 들어가는지 확인하는 것이 T검정의 최종목적이다.

(<https://m.blog.naver.com/lavenda999/221396494475>)

다음은 독립 이표본 검정을 사용하는 경우의 예시 이다.

Ex) 1반, 2반 두 집단의 영어 성적을 비교하고 싶다.

반이 2개 -> 집단이 2개

각 반은 독립적

* 독립 표본 T검정, Two sample t test를 사용한다.

Ex) 동일한 제품을 생산하는 A공장의 일일 생산량과 B공장의 일일 생산량의 차이가

통계적으로 의미있는 차이인지 비교하고 싶다.

공장 2개 -> 집단이 2개

각 공장은 독립적

* 독립 표본 T검정, Two sample t test를 사용한다.

Ex) 우리나라 30대 남성과 여성의 평균 신장은 통계적으로 유의한 차이가 있는지

비교하고 싶다.

남성, 여성 -> 집단이 2개

비교 집단은 독립적

* 독립 표본 T검정, Two sample t test를 사용한다.

(<https://blog.naver.com/statsol/221320719271>)

독립표본 T검정 사용의 조건은

1. 독립성

(두 변수는 서로 독립된 값이어야 한다.)

1. 정규성

(각 변수들은 정규분포여야 한다.

중심극한의 정리에 의해 각 변수의 표본이 30개 이상이면 정규분포로 볼 수 있다.)

1. 등분산성

(통계적 모형에서 가정하는 관측값의 오차항이 모든 관측값에 대해 같은 분산을 가지는 것을 말한다.)

3가지 가정이 만족되어야 한다.

분석을 하고자 하는 집단에서 관측되거나 측정된 수치는 다양한 형태로 나타날 수 있다.

키와 몸무게, 혈압, 콜레스테롤 수치와 같은 값들은 두 숫자 사이에 끊임없이 다른 숫자가 나타날 수 있는 연속형 수치이며 또한 동일 간격을 가진다.

예를 들어 160cm와 170cm에는 10cm의 간격이 존재한다.

170cm와 180cm 사이에도 10cm란 간격이 존재하는데, 이들 두 가지 10cm의 간격은 서로 동일한 의미를 가진다는 것이다. 한편, 20cm은 10cm에 비해 2배 크다고 할 수 있는데, 이러한 특성을 가진 자료를 비 척도(ration scale)자료 라고도 한다.

IQ나 온도와 같은 수치도 동일 간격을 가지는 연속형 수치이다.

하지만 위의 예시와는 달리 한 숫자를 다른 숫자에 비해 몇 배 더 크다라고 할 수는 없다.

IQ 200이 IQ100보다 2배 더 똑똑하다고 할 수 없다. 섭씨 40도 온도가 섭씨 20도 온도보다

2배 더 덥다고 할 수도 없다. 이러한 특성을 가진 자료를 간격 척도(interval scale)자료 하고 한다.

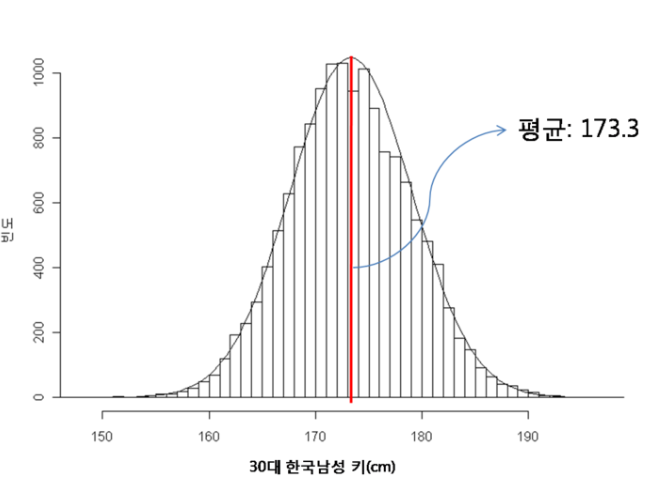
이렇게 자료가 동일 간격을 가진 연속형 수치인 비 척도 또는 간격 척도인 경우에만 독립 표본 T검정 방법을 사용할 수 있다.

정규성 이란 우리가 수집한 데이터가 특정 값에 편중됨 없이 적절히 잘 수집되었나를 확인하는 과정이다. 정규성을 나타내는 자료 분포를 통계학 용어로 정규 분포(normal distribution)라고 한다.

아래 그림은 15000명 30대 한국남성의 키 분포를 나타낸 그림이다.

X축은 센티미터로 표시된 키 범위를 나타내며 Y축은 해당 키 범위에 속한 사람 수를 나타낸다.

그림에서는 측정된 키가 평균 173.3cm를 기준으로 하여 좌, 우가 거의 대칭구조를 보이고 있는데, 이런 형태의 자료 분포를 정규 분포라고 한다.



독립 표본 T 검정에서 많은 사람들이 헷갈려 하는 것이 자료의 정규성 문제이다.

일반적으로, “각 집단에서 30개 이상의 측정값을 확보한 경우에는 자료가 정규성을 가진다.”라는 말을 많이 듣게 된다. 여기서 중요한 사실은, 위에서 언급한 자료의 정규성이 각 집단에서 확보한 30개 측정값 자체의 정규성을 의미하는 것이 하니라 30개 측정값의 평균이 가지는 정규성이라는 것이다.

즉, 독립 표본 T 검정에서는 두 집단에서 측정한 수치 하나 하나를 비교하는 것이 아니고 각 집단의 평균을 비교하는 것이기 때문에 자료의 정규성 이라는 것 또한 각 평균이 가지는 정규성을 의미하게 된다.

표본 집단(sample)의 표본 수가 30개 이상일 경우 표본 평균(sample mean)은 모집단(population)의 자료 분포와는 상관없이 정규 분포를 따른다(정규 분포에 가까워진다)라는 것은 중심 극한 정리(central limit theorem)를 통해 증명된 사실이다.

측정값이 30개 미만인 자료인 경우에는 30개 미만인 측정값 자체 만으로 정규성 검증을 하게 된다. 이때 표본 집단 자료가 정규성을 만족하게 되면 표본 평균도 정규성을 만족하기 때문에 독립 표본 T 검정 방법을 사용할 수 가 있다.

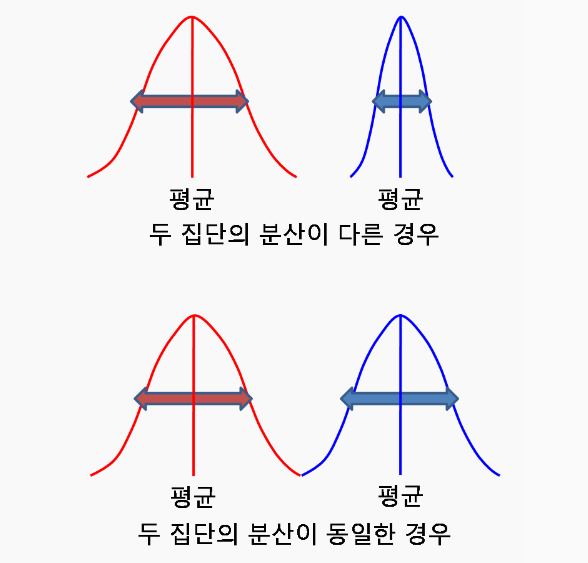
만일 30개 미만인 표본 집단 자체가 정규성을 만족하지 못할 경우에는 순위 척도 자료인 경우에 사용하는 비모수적 검정 방법인 Mann-whitney U test를 사용할 수 있다.

자료의 정규성을 검증하는 방법으로는 Kolmogorov-Smirnov-test와 Shapiro-Wilk test방법이 있다.

한편, 독립 표본 T 겁정의 경우 Student’s T test와 Welch’s T test 두 가지가 있다.

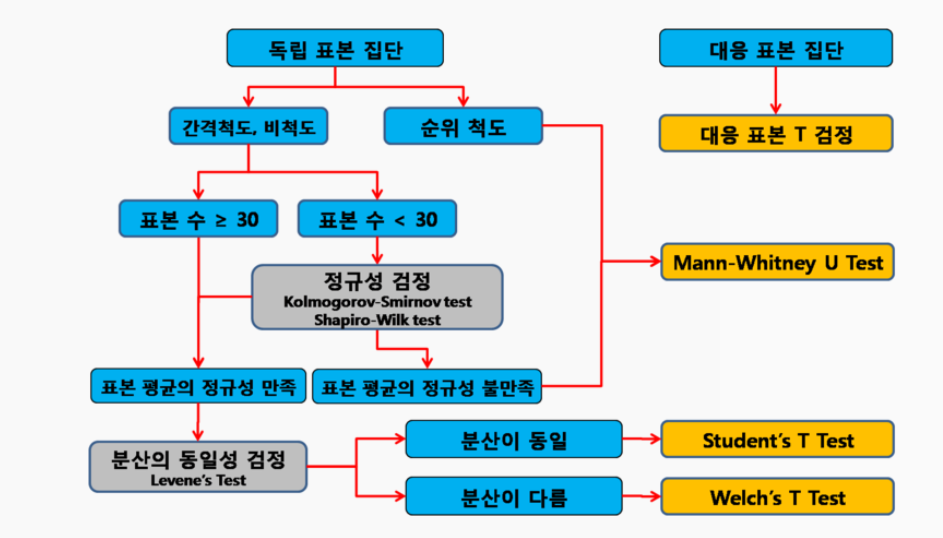
두 집단 각각에서 추정된 분산(variance)이 동일한지 혹은 다른지에 따라 사용 경우가 다르다.

만인 분산이 동일하다고 판단되면 Student’s T test를, 그렇지 않을 때는 Welch’s T test 사용한다.



두 집단 자료의 분산이 동일한지 여부는 Levene’s test방법을 이용해 분석할 수 있다.

두 집단간 차이 분석을 하는 방법을 요약하면 아래 그림과 같다.



(<https://blog.naver.com/istech7/50150957927>)

// 기본적으로 알아두어야 하는 사항

간단하게 영가설, 대립가설, 유의수준, 유의확률에 대해 정리해보았다.

H0 영가설(귀무가설) : 두 집단 평균은 차이가 없다.

H1 대립가설(연구가설) : 두 집단 평균은 차이가 있다.

통계를 돌렸을 때

H0 영가설, 귀무가설 -> ‘차이 없다’고 세운 가설 [기각]하고

H1 대립가설, 내주장 가설 -> ‘차이 있다’고 세운 가설을 [채택]해야 연구 성공이다.

통계의 결과가 숫자로 나오는데 이 숫자를 유의수준, 유의확률(p-value)라고 한다.

만약, H0 영가설이 참인데 기각해 버리는 실수를 저지를 확률,

내 실험결과가 맞다고, ‘오류를 범할 확률’ 그것이 p-value 유의확률이다.

유의확률은 실수를 저지를 확률이므로 적어야 좋은 것이다.

어느정도면 영가설을 기각하고, 대립가설인 연구결과를 채택하고 받아들일 것이냐의

수준은 연구자가 정한다. 그 수준이 유의수준, significance level이다.

유의수준을 a라고 쓰는데, 0.05 아니면 0.01 아니면 0.001 이렇게 사용하며

이것은 5%, 1%, 0.1%를 나타낸다.

보통 p앞에 별을 붙여준다. 0.05는 별 하나, 0.01는 별 두 개, 0.001 보다 작은 p-value는 별 세 개,

\*p<.05

\*\*p<.01

\*\*\*p<.001

이렇게 나타낸다.

2) 데이터 분석

R script:

# mpg 데이터 불러와 class, cty 변수만 남긴 뒤 class 변수가 "compact"인 자동차와 "suv"인 자동차를 추출한다.

mpg <- as.data.frame(ggplot2::mpg)

# dplyr패키지는 데이터 전처리 작업에 가장 많이 사용되는 패키지

# filter(), select(), arrange() 등등의 함수가 있다.

# filter() 행 추출

# select() 열(변수) 추출

# %>% 기호는 함수를 나열하는 방식(파이프 연산자), 단축키 Ctrl + Shift + M

library(dplyr)

mpg\_diff <- mpg %>%

select(class, cty) %>%

filter(class %in% c("compact", "suv"))

# 앞에서부터 6행까지 출력

head(mpg\_diff)

# 빈도표로 compact, suv 수 알아보기

table(mpg\_diff$class)

# t.test()는 R에 내장된 함수이다.

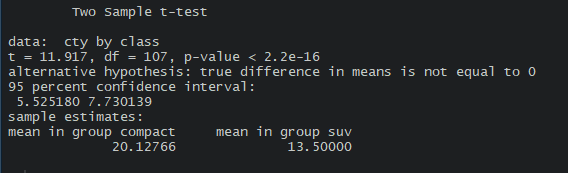
# 추출한 mpg\_diff 데이터 지정, ~기호를 이용하여 비교할 값인 cty(도시 연비)변수와 비교할 집단인 class(자동차 종류) 변수를 지정한다.

# t 검정은 비교하는 집단의 분산(값이 퍼져 있는 정도)이 같은지 여부에 따라 적용하는 공식이 다르다.

# 여기서는 집단 간 분산이 같다고 가정하고 var.equal 에 T를 지정하였다.

t.test(data = mpg\_diff, cty ~ class, var.equal = T)

결과값:



p-value 는 유의 확률을 의미한다.

p-value < 2.2e-16 은 2.2앞에 0이 16개 있는 값(2.2\*10의-16승)보다 작다는 의미이다.

p-value가 0.05보다 작기 때문에 이 분석 결과는 ‘compact’와 ‘suv’간 평균 도시 연비 차이가 통계적으로 유의하다고 해석할 수 있다.

Sample estimates를 보면 각 집단의 cty 평균이 나타나 있다. ‘compact’는 20이고, ‘suv’는 13이므로 ‘suv’보다 ‘compact’의 도시 연비가 더 높다고 할 수 있다.

# 일반 휘발유(Regular)를 사용하는 자동차와 고급 휘발유(Premium)를 사용하는

# 자동차 간 도시 연비 차이가 통계적으로 유의한지 알아보자.

# r : regular, p : premium

mpg\_diff2 <- mpg %>%

select(fl, cty) %>%

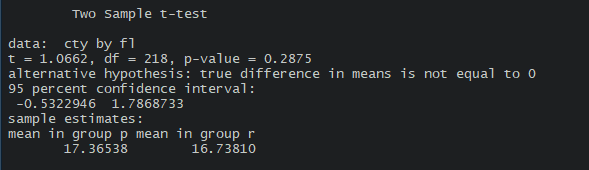
filter(fl %in% c("r","p"))

head(mpg\_diff2)

table(mpg\_diff2$fl)

t.test(data = mpg\_diff2, cty ~ fl, var.equal = T)

결과값:



p-value가 0.05보다 큰 0.2875이다. 실제로는 차이가 없는데 우연에 의해 이런 차이가 관찰된 확률이 28.75%라는 의미이다. 따라서 일반 휘발유와 고급 휘발유를 사용하는 자동차 간 도시 연비 차이가 통계적으로 유의하지 않다고 결론 내릴 수 있다.

Sample estimates를 보면 고급 휘발유 자동차의 도시 연비 평균이 0.6 정도 높지만 이런 정도의 차이는 우연히 발생했을 가능성이 크다고 해석하는 것이다.

(쉽게 배우는 R데이터 분석 책 )

3. 데이터 출처 : R studio ggplot2패키지 mpg데이터

mpg(미국 환경 보호국)에서 공개한 1999~2008년 사이 미국에 출시된 234종의 연비 관련 정보

4 외부 데이터 첨부 : R studio 데이터 사용함. 외부 데이터 사용하지 않음.

-이상-