가설검정(python)

pyk

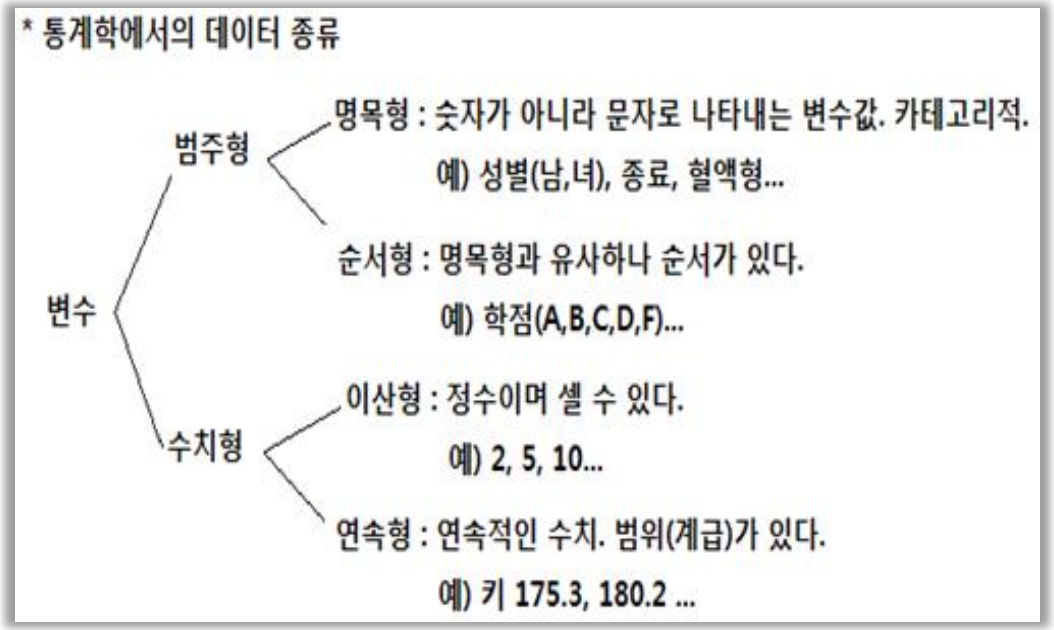
* 척도(Scale) :

- 자료가 수집될 때 관찰된 현상에 하나의 값을 할당하기 위해 사용되는 측정의 수준

- 척도(자료가 수집될 때 관찰된 현상에 따라 하나의 값을 할당하기 위해 사용되는 측정의 수준)에 따른 분류

  : 범주형(정성적) - 수량화가 불가

  : 수치형(정량적) - 수량화가 가능



또는 아래와 같이 구분하여 불리기도 한다.

- 명목척도 (nominal scale) – 척도의 명칭만 의미 있음
  (예) 결혼 상태에 대한 코드: { 미혼=1, 기혼=2, 돌싱=3, 사별=4}
  성별 : 남 = 1, 여 = 2, 혈액형

- 서열척도 (순서척도, ordinal scale) – 명칭 및 순서가 의미를 지님
  (예) 성적 등급 - {poor=1 , fair=2 , good=3 , very good=4}
  건강상태 - {나쁨=1, 보통=2, 양호=3}

- 등간척도 (간격척도, interval scale) – 명칭, 순서 및 간격이 의미를 지님
  연속형 자료이나 절대적 원점이 없다.
  (예) 온도, 물가지수 생산지수 등

- 비율척도 (ratio scale) – 명칭, 순서, 간격 및 배율 모두 의미를 지님
  이들 척도의 경우 이른바 "절대적 원점(absolute zero point)"이 정의됨
  (예) 키, 몸무게, 재산, 가격, 소득, 광고비, 판매량, 매출액 등

척도에 따른 데이터 분석방법

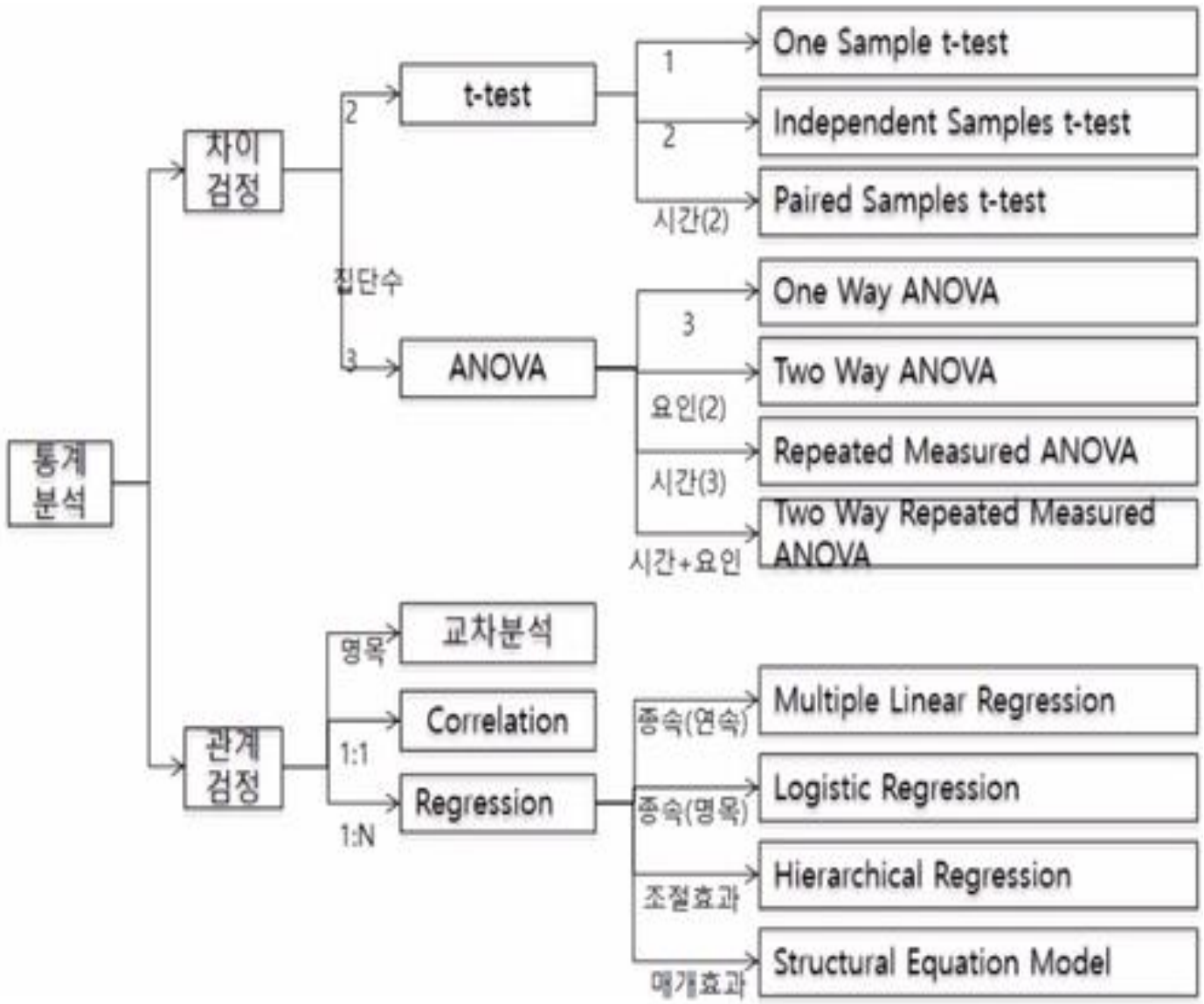
데이터 분석(추론 및 검정)을 하는 경우에 있어 종속변수에 영향을 주는 독립변수(설명변수, 원인변수)와 영향을 받는 종속변수(반응변수. 결과변수)가 있다.

이러한 독립변수와 종속변수는 원인과 결과의 관계를 갖는다.

독립변수 x (영향 줌)	종속변수 y (영향 받음)	분석 방법
범주형	범주형	카이제곱 검정
범주형	연속형	T검정 (범주형 값 2개 : 집단 2개 이하), ANOVA (범주형 값 3개 :집단 3개 이상)
연속형	범주형	로지스틱 회귀분석
연속형	연속형	회귀분석, 구조 방정식

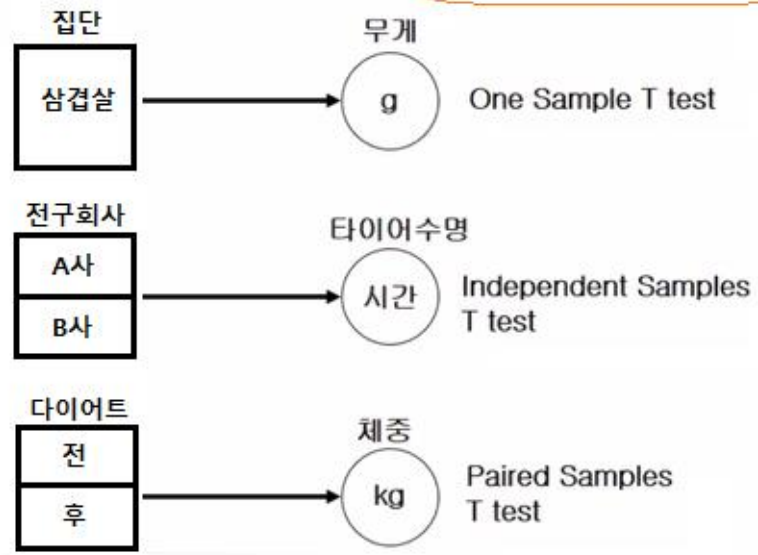
통계분석방법

- 평균차이검정
  - 집단간 평균차이를 검정하는 방법
  - 평균검정(T-test), 분산분석(ANOVA), ANCOVA, MANOVA
- 관계검정
  - 변수와 변수의 관계를 검정
  - 상관분석, 회귀분석, 교차분석, 정분상관분석, 판별분석, 로지스틱회귀분석
- 신뢰도와 타당도
  - 신뢰도분석, 요인분석
- 기타
  - 군집분석, 다차원척도법, 생존분석, 데이터마이닝 기법등

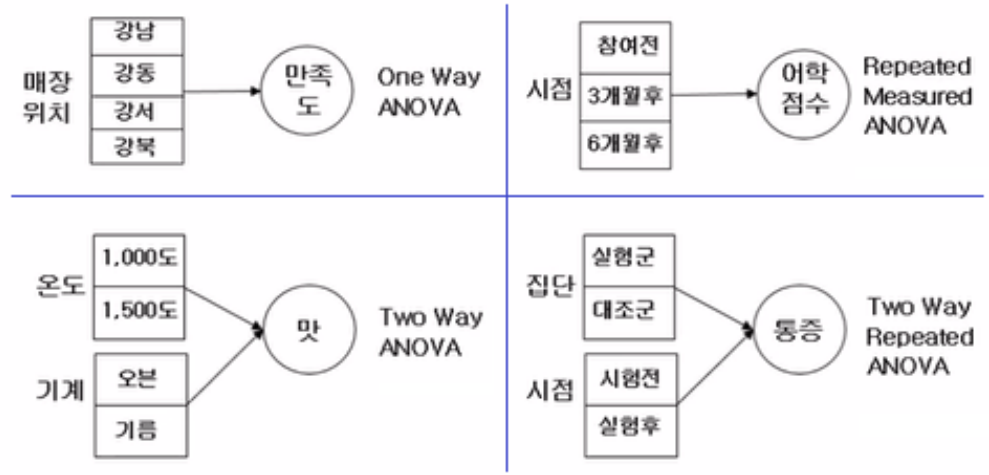


T-test (T 차이 검정)

X 명목변수(Categorical: C)
Y 연속변수(Metric: M)

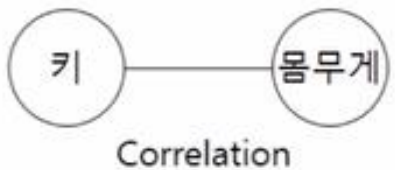


ANOVA



- 관계검정 : **변수 사이의 관계를 알아보는 검정방법**
  관계검정은 독립, 종속변수의 type이 같다.
  - 변수와 변수의 관계를 검정
  - 상관분석(Correlation Test): **연속형(X) + 연속형(Y)**
  - 회귀분석(Regression): **연속형(X) + 연속형(Y)**
  - 교차분석(test): **범주형(X) + 범주형(Y)**
    - 2개의 질적변수

- 상관분석(Correlation Test)
  - 상관관계: 두 변수가 서로 동등한 입장에서 관계를 분석
    예) 몸무게가 많이 나가면 허리둘레도 크고, 반대로 허리둘레가 크면 몸무게도 많이 나감
  - 편상관분석(Partial Correlation) : 중간에 다른 변수의 영향력이 있을 때 이를 통제

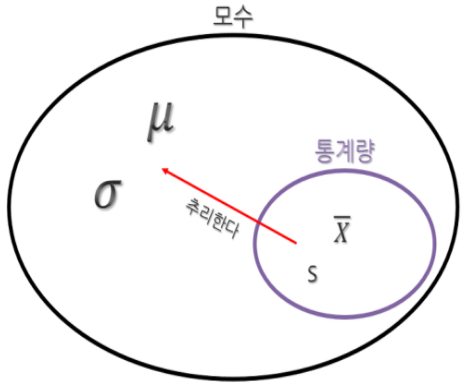


- 회귀분석(Regression)
  - 인과관계: 하나의 변수가 원인이 되어 다른 변수(들)에 영향을 미치는 관계

통계적 추정이란?

통계에서 추정(estimation) 추정은 무엇이고 왜 하는 것일까?

보통 어느 모집단의 특징을 알고 싶어서 조사를 진행하는 경우가 많다. 하지만 현실에서는 시간과 비용의 제약이 있어 모 집단의 전부를 조사하기란 매우 어렵다. 그래서 집단의 특징을 파악하기 위해 모집단 전체를 조사하는 것이 아니라, 표본자료를 추출해서 조사하게 된다.



이때 알고자 하는 모집단의 특징을 숫자로 나타낸 것을 "모수(μ, σ)"라 하고, 표본 데이터를 "통계량($\bar{x}$, s)"이라고 하며, 표본의 통계량을 통해 모수가 "이러할 것이다."라 추리하는 것을 추정이라고 한다. 그런데 추정은 일부의 데이터만 가지고 전체를 추리하는 것이기 때문에, 100% 정답이 아니라 다소의 오차가 발생하게 된다는 문제가 있다.

하지만 아무리 오차가 있다 해도 현실에서는 시간과 비용의 제약으로 어느 정도의 오차를 인정한 체 표본만으로 조사를 진행하지, 모집단 전체를 조사하는 경우는 거의 없다. 그래서 통계에서는 표본의 수치만 가지고 모집단의 특징을 추리하는 추정이 일반화되어 있다.

한편 추정은 크게 점추정과 구간추정으로 나눌 수 있다. 표본의 통계량을 가지고 모집단의 모수를 추리하는 것을 추정이라고 했다. 예를 들어 한국 성인 남자의 평균 키(모집단)를 파악하기 위해, 성인 남자 1000명을 표본으로 뽑아 키를 조사하였더니, 평균이 172.34cm가 나왔다고 할 때, 172.34cm처럼 하나의 수치, 즉 하나의 점으로 값을 표현하는 것이 **점추정(point estimation)**이다.

그런데 1000명을 대상으로 나온 수치가 172.34cm(통계량)라고는 하지만, 정말로 한국 성인 남자의 평균 키(모수)가 172.34cm일까? 아쉽지만 그럴 확률은 거의 없다. 값을 신뢰하기에 표본의 수가 너무 적을 뿐 더러, 표본에는 항상 오차가 동반된다. 거기에 소수점 단위를 더욱 세분화하면, 점추정치 172.34cm가 모수와 같을 확률은 거의 제로에 가깝다. 이렇게 점추정치는 그 특성상 값을 신뢰하기가 어렵다.

구간추정(interval estimation)

점추정치의 한계를 극복하기 위해 점추정치를 기준으로 일정구간을 설정하는 방법이 있다. 점추정치가 172.34cm가 라면, 이를 기준으로 ±5를 해서 일정구간(167.34cm~177.34cm)을 만든다. 그러면 이 구간 안에 모수가 포함될 확률은 자연스럽게 높아진다. 이렇게 추정치의 신뢰도를 높이기 위해 점추정치를 중심으로 일정 구간을 만드는 것이 구간추정(interval estimation)이다.

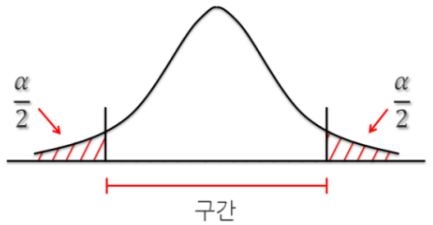
그런데 구간추정이라고 100% 신뢰할 수 있는 것은 아니다. 경우에 따라서는 구간추정치 안에 모수가 포함되지 않을 가능성도 항상 존재한다. 그리고 구간추정은 점추정에 비해 신뢰도가 높다고 할 수 있지만, 점추정이 전혀 필요 없는 것은 아니다. 왜냐하면 점추정을 기준으로 구간추정을 하기 때문이다.

신뢰구간이란?

구간추정을 할 때는 과연 어느 정도로 구간을 만드는가? 라는 문제가 있다. 예를 들어 한국 성인 남자의 평균 키를 150cm~190cm로 구간추정 했다고 하자. 그런데 이 구간은 너무 넓어서, 평균 키(모수)가 150cm~190cm 구간 안에 들어가는 것은 당연하다. 하지만 신뢰하기에는 구간이 너무 넓다.

신뢰구간은 되도록 좁을수록 좋은데, 그렇다고 너무 좁으면 모수가 포함되지 않을 확률이 높아지기에 너무 좁게 설정할 수도 없다. 그래서 구간추정으로 구간을 만들 때는 적정한 구간을 만들 필요가 있는데, 나름의 기준을 통해 신뢰할 수 있는 구간을 만든 것이 신뢰구간(confidence interval)이다.

그러나 아무리 신뢰할 수 있는 구간이라도, 모수가 신뢰구간 안에 포함되지 않을 확률은 항상 존재하는데, 이 확률을 α(알파)라고 한다. 신뢰구간은 양쪽(왼쪽과 오른쪽)을 다루어야 하므로, α가 둘로 나뉘어서 α/2가 된다. 그래서 신뢰구간을 추정할 때는 α/2가 많이 나온다. 정규분포 그래프로 확인할 수 있다.

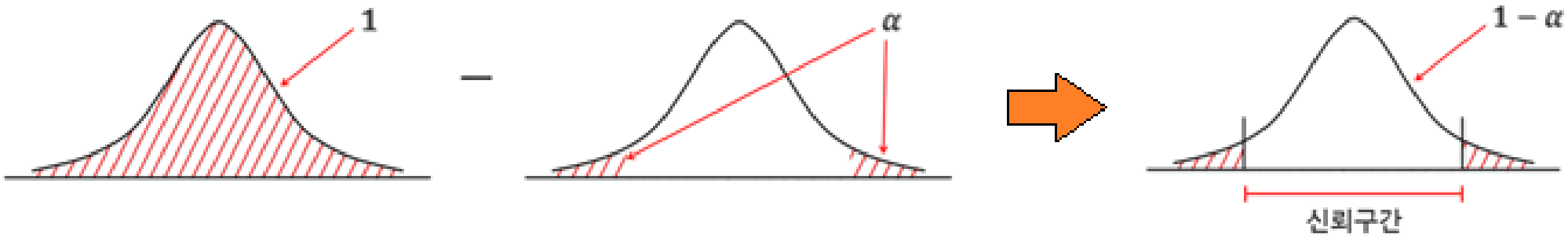


신뢰수준

확률의 총합은 1 (100%) 이다. 그래서 모수가 구간에 포함되지 않을 확률이 α이므로, 모수가 구간에 포함될 확률은 1-α가 된다.

이 1-α를 보통 신뢰수준(신뢰계수)이라고 하는데, 보통 90%, 95%, 99%의 확률을 많이 사용한다. 그리고 이 90%, 95%, 99%의 신뢰수준을 통해 설정된 구간이 신뢰구간이다.

이렇게 신뢰구간을 설정할 때는 임의대로 대충 잡는 것이 아니라, 신뢰수준(1-α)을 기반으로 해서 신뢰구간을 설정한다.



신뢰구간은 크게 "모평균의 신뢰구간"과 "모분산의 신뢰구간" 그리고 "모비율의 신뢰구간"을 많이 구하는데, 이러한 신뢰구간을 추정할 때는 각각에 맞는 확률분포를 사용해서 구한다.

통계적 가설검정이란?

가설검정이라면 모수(母數 , Population Parameter)에 대한 주장을 가설로 정립한 것으로 가설이 맞는 지를 자료를 통해 판단할 수 있다. 자료에서 필요한 통계량을 계산한 후에 어떠한 값이 나온 확률을 계산한 후 판정하게 된다.
세상에는 철학, 역사, 문학, 경영, 경제, 사회, 과학, 수학 등 여러 분야에 걸쳐서 많은 수의 이론이나 관념들이 있다. 하지만 이러한 이론이나 관념들은 그 자체로써 신뢰하기 보다는, "아마도 이럴 것이다."라는 하나의 가설에 불과하다. 정설이 아니라 하나의 가설이기에, 이러한 이론이나 관념에는 항상 불완전함(불신)이 내포되어 있다.

이렇듯 불완전하기에 새로운 가설이 생겨나고, 현재의 가설은 새로운 가설에 의해 대체되기도 한다.

신화적 상상력(가설) → 천동설(가설) → 지동설(가설) → ?(가설)

그런데 새로운 가설이 나왔다고 해서, 현재의 가설보다 더 합리적이고 신뢰할 수 있는 것은 아니다. 오히려 새로운 가설보다 현재의 가설이 더 정확한 경우도 있다. 그래서 새로운 가설이 나오면, 현재의 가설을 폐지하고 바로 새로운 가설로 대체하는 것이 아니라, 두 개의 가설 중 어떤 가설이 더 정확하고 신뢰성이 있는지를 판단하는데, 이 판단하는 과정이 가설검정이다.

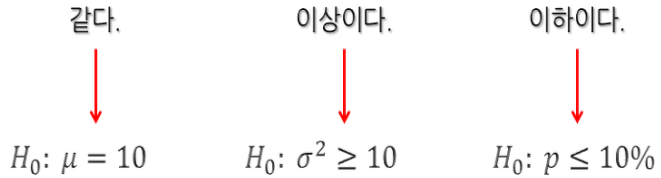
현재의 가설 VS 새로운 가설

통계에서는 현재의 가설을 귀무가설(H0)이라고 하고, 새로운 가설을 대립가설(H1)이라고 하는데, 가설검정은 통계관련 수식을 활용해서 이 귀무가설과 대립가설 중 어느 가설이 더 타당한지를 판단한다.

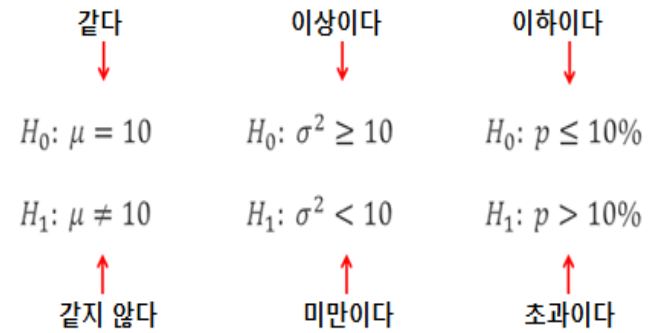
귀무가설과 대립가설 설정하는 방법

가설검정 절차 중 가장 먼저 하는 것이 귀무가설과 대립가설 설정으로, 두 개의 가설은 정반대로 설정되어야 한다. 왜냐하면 가설검정은 귀무가설과 대립가설 중 어느 것이 더 타당한지를 판단하고, 하나의 가설을 양자택일하는 것인데, 하나를 선택하기 위해서는 두 개의 가설이 중복됨이 없이 정반대여야 가능하다. 그래서 정반대로 설정한다.

귀무가설(영가설)은 "~와 같다."와 "~이상이다." 그리고 "~이하이다." 와 같이 3가지 유형이 있다. 예를 들면 "평균은 10과 같다." "분산은 10 이상이다." "비율은 10% 이하이다."와 같이 표현할 수가 있다. 그래므로 =와 ≥와 ≤의 부등식을 사용해서 설정한다. 귀무가설을 구체적으로 표현하면 옆의 그림과 같다.



대립가설은 귀무가설과 정반대로 설정해야 하므로, ≠와 < 와 >의 부등식을 사용해서 설정한다. 위의 예를 활용하면, "평균은 10과 같지 않다." "분산은 10 미만이다." "비율은 10% 초과이다."와 같이 풀어낼 수가 있다. 옆의 그림과 같이 기호로 표현할 수 있다.



참고로 귀무가설과 대립가설을 설정할 때, 보통 표본에서 사용하는 기호인 와 s2과 $\hat{p}$ $\bar{X}$ 은 사용하지 않는다. 왜냐하면 가설은 "모집단의 모수가 이럴 것이다."라고 표현한 말이기 때문에, 당연히 모수인 μ와 σ2과 p만 사용해서 가설을 표현한다. 표본의 통계량은 단지 가설을 세우기 위한 하나의 재료일 뿐, 귀무가설과 대립가설 설정에는 들어가지 않는다.

* 귀무 / 대립가설 설정 연습 *

연습1)
개와 고양이의 평균수명을 조사하였다. 두 집단의 평균수명이 차이가 있는지를 파악하기 위해 가설검정을 하려 한다. 귀무가설과 대립가설을 설정하라.

연습2)
새우깡 과자를 생산하는 기계1과 기계2가 있다. 기계1에서 생산한 제품의 분산이 큰 것으로 알려져 있다. 과연 그런지 검정하려고 하는데, 여기에 적당한 귀무가설과 대립가설을 설정하라.

연습3)
A제품과 B제품이 있는데, A제품의 품질불량에 대한 항의전화가 많이 온다고 한다. 이런 이유로 A제품의 불량률이 B제품의 불량률보다 더 클 것이라는 얘기가 나오고 있다. 실제로 그런지를 알아보기 위해 가설검정을 하려 할 때, 귀무가설과 대립가설을 설정하라.

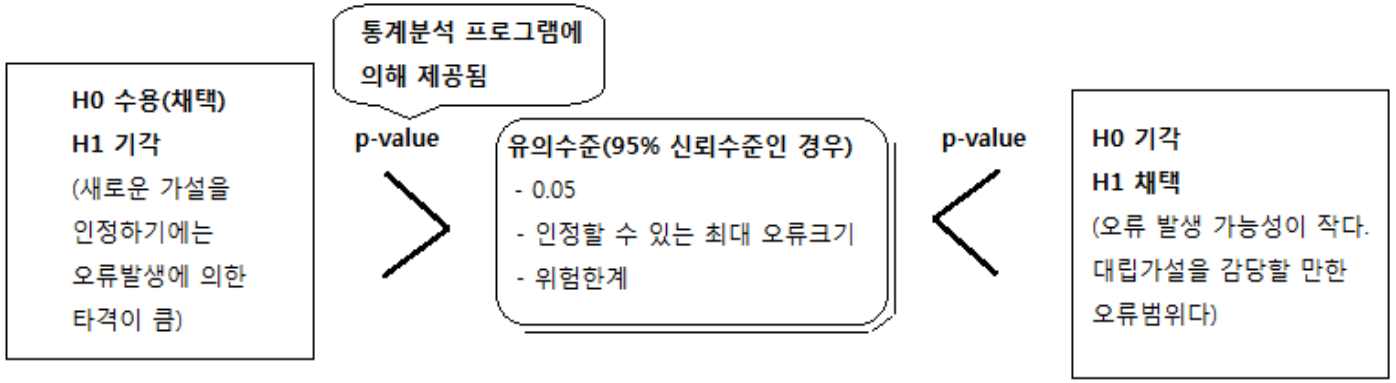
유의수준 / 유의확률

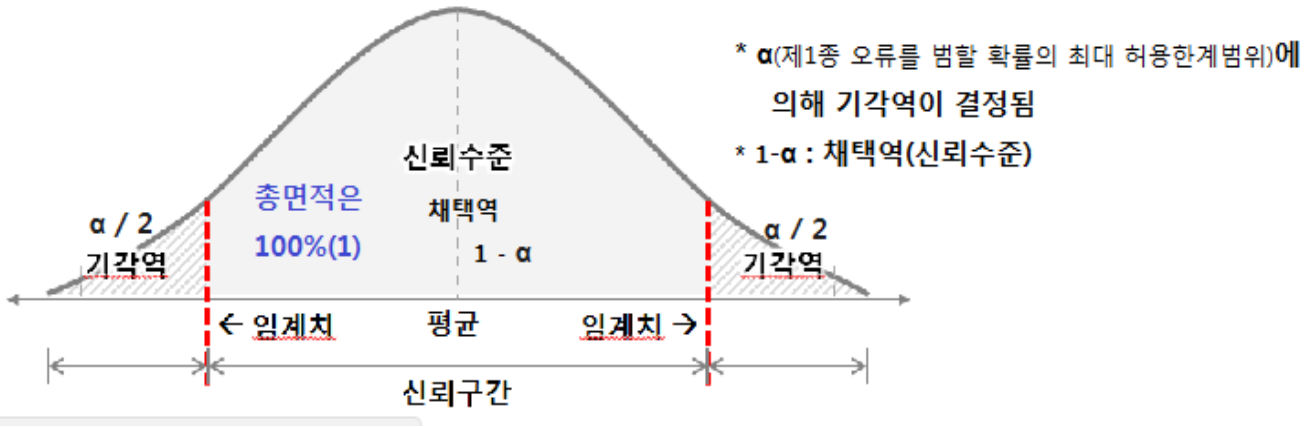
통계처리를 할 때 대개는 모집단의 분산이나 평균을 알기가 어렵다. 모집단 자체를 전수조사하기가 힘들기 때문이기도 하고, 모집단의 정확한 범위를 알지 못 하는 경우도 있다. 그래서 실제 모집단에서 표본 몇 십 개 또는 몇 백 개를 추출해서 그것의 분산(표본분산)이나 평균(표본평균)을 사용해야 하는 경우가 대부분이다.

예) 두 개의 모집단에 대한 모평균의 동일 여부를 검정할 때, 아래와 같이 가설을 세우게 된다.
- 귀무가설 H0 : 두 개의 모평균이 같다.
- 대립가설 H1 : 두 개의 모평균은 같지 않다.

유의수준(significance level)은 보통 1%, 5%, 10% 세 개를 주로 사용하는데, 즉, 0.01, 0.05, 0.1.... 그 중에서 1%와 5%를 많이 사용한다. 유의수준 0.05라 함은 예를 들어 두 개 집단의 모평균은 실제 같은데도 잘못 판단 해서 귀무가설을 기각하게 될 확률(모평균이 같지 않은 것으로 판단할 확률)을 의미한다. "1종 오류"를 범할 확률이다.
다시 말하면, **5%(0.05)의 유의성이란 테스트 결과가 "사실이 아닐 확률"이 5%**라는 뜻. 또는 "사실일 확률"이 95% 라는 얘기다. 즉, 이와 같은 테스트 방법을 100번 사용할 때 95번 정도만 맞게 검정한다는 뜻이다.

유의확률(p-value)은 귀무가설을 기각할 수 있는 최소한의 확률을 말한다. 가령 유의수준 5%에서 유의확률이 0.009로 도출되었다고 해보자. 유의확률은 유의수준을 보다 정확히 계산한 것으로 귀무가설을 잘 못 기각할 확률은 0.9% 밖에 안 된다는 것이다. "1% 이내냐 5% 이내냐" 라는 식으로 대충 얘기하는 것이 유의수준이고, 테스트했더니 잘못될 확률이 정확하게 얼마냐 하고 계산한 것이 유의확률이다. 여기서는 0.009(0.9%)니까, 5%의 유의수준에서 귀무가설을 기각할 수 있게 된다.
즉, 두 모집단의 모평균이 서로 같다는 귀무가설을 잘 못 기각할 확률이 5% 이하(정확히 말하면 0.9%)라는 뜻이다. 그러니까, 두 집단 간에는 통계적으로 유의미한 차이가 있다는 말이다. 그래서 "유의"라는 말을 쓰고 있다.





p값(유의 확률)의 정의

p값을 사용하여 결과가 통계적으로 유의한지를 확인할 수 있다. p값은 귀무가설을 기각하거나 채택하는 가설 검정에서 주로 사용된다. 가설 검정을 수행하는 경우 결과 도출시에 중요한 부분은 p값이다. P값의 범위는 0 ~ 1 이다. p값은 귀무가설에 반하는 증거를 측정하는 확률이다. p값이 작을수록 귀무가설에 반하는 더 강력한 증거가 된다.

p값을 α(유의수준)에 비교하여 귀무가설을 기각해야 하는지 여부를 결정할 수 있다. p값이 α보다 작거나 같으면 H0를 기각하고 p값이 α보다 크면 H0를 기각할 수 없다.

신뢰도 95%라면 5%(0.05) 값이 α에 많이 사용되므로, p값이 0.05보다 작거나 같으면 H0를 기각한다.

‘p값이 5% 수준에서 유의하다’는 뜻은?

예를 들어 기존에 ‘한국남자 키의 평균이 173으로 알려져 있다’면 이것이 귀무가설이고, ‘이보다 커졌다’라는 반론은 대립가설이다.

통계분석을 실시할 때 유의수준(알파)과 유의확률(p값)을 비교하여 귀무가설의 기각 여부를 판단한다.

일반적인 유의수준은 0.05(5%)로 정한다. 유의수준이란 귀무가설이 사실일 때 귀무가설을 기각할 오류의 최대 허용범위다. 5% 정도는 잘 못 판단할 수 있으며 이를 감수한다는 기준 값이다. p값이 유의수준 5%에서 유의하다는 것은 p값이 유의수준 0.05보다 더 작게 나왔다는 의미다. 따라서 귀무가설을 지지하는 정도가 우리가 정해놓은 기준보다 작아 '대립가설을 채택하겠다'라는 의미다.

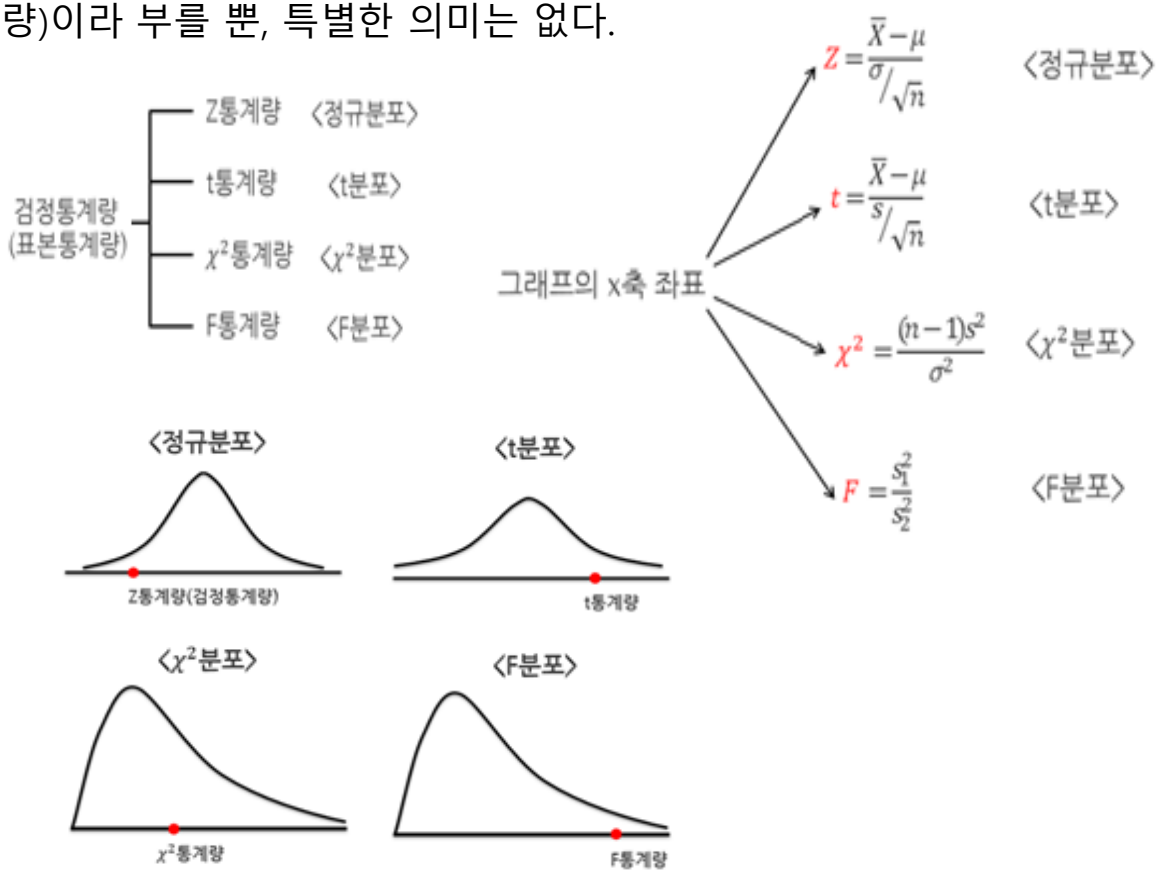
검정통계량이란?

어느 가설이 맞는가를 판정할 때 기준이 되는 값이다. 예를 들어 '평균차이가 크다. 표본비율의 값이 크다' 등. 이는 상수가 아니라 확률변수다. 그러므로 언제든지 틀릴 가능성은 존재한다. 가설검정은 "모집단의 모수가 이럴 것이다."라는 가설을 다루며 p값을 사용해서 귀무-대립가설을 설정한다.

가설이 타당한지를 파악하기 위해, 계산을 할 때는 모수를 사용할 수가 없다. 왜냐하면 시간과 비용이 너무 많이 들어 현실적으로 모집단 전체를 조사할 수는 없기 때문이다. 그래서 통계에서는 표본통계량으로 계산을 하곤 하는데, 이 표본통계량을 가설검정에서는 검정통계량이라고 부른다. 이는 가설검정에서 사용하는 통계량이기에 검정통계량(=표본통계량)이라 부를 뿐, 특별한 의미는 없다.

가설검정은 대충 하는 것이 아니라 확률분포를 활용하는데, 신뢰구간 추정과 마찬가지로 정규분포, t분포, χ2분포, F분포를 이용한다.
그래서 검정통계량도 확률분포에 따라 Z통계량, t통계량, χ2통계량, F통계량으로 세분화할 수 있다.

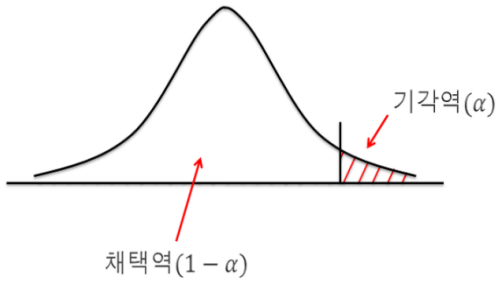
관심있게 볼 사항은 검정통계량으로 확률을 구하는 것이 아니라, 확률분포 그래프의 x축 좌표를 구한다는 점이다.
그래서 검정통계량의 공식을 보면, 그래프의 x축 좌표인 Z값, t값, χ2값, F값 구하는 공식인 것을 알 수 있다.



기각역이란?

가설검정은 귀무가설과 대립가설 중에서 하나를 선택한다. 가설검정은 나름의 기준을 통해 채택과 탈락 여부를 결정하는데, 결정 결과가 100% 정답이 아니라 어느 정도의 오차는 있을 수 있다. 그래서 가설검정도 틀릴 확률은 항상 존재한다. 이러한 오차에 의해 1종 오류가 생길 수 있는데, 이 확률을 α(유의수준)라고 한다. 그렇다면 "귀무가설을 채택할 확률"은 1-α (확률의 총합은 1)가 된다.

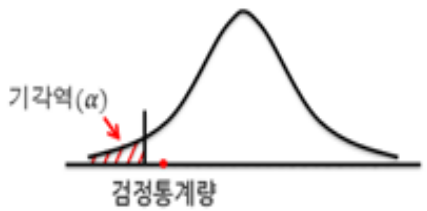
확률 1-α는 귀무가설을 채택하게 됨으로, 이 영역을 "채택역"이라 하고 α의 영역을 "기각역"이라 부른다.



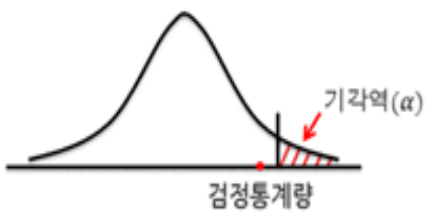
<귀무가설 채택>

<귀무가설 탈락>

채택역과 기각역으로 귀무가설의 채택과 기각 여부를 판단하는데, 그림을 보면 x축 값의 위치가 중요하므로 임계치를 구하기 위해 검정통계량을 활용한다. 검정통계량의 결과가 채택역 안에 위치하면 귀무가설이 채택되고, 아니면 귀무가설이 기각(탈락)된다.



<좌측검정>



<우측검정>

<양측검정>

모평균의 가설검정(σ를 아는 경우)을 해보자.

H_0: μ=300
H_1: $\mu \neq 300$

모평균의 가설검정은 "μ가 이럴 것이다. 아니다"라고 대립한 두 개의 가설 중, 어느 가설이 더 타당한지를 판단하고, 두 가설 중 하나가 옳다고 판단하는 것이다. 그런데 가설검정은 판단만 하면 안 되고, 결론까지 내줘야 한다. 문제를 통해 가설검정을 진행해 보자.

* A사 건전지 1개의 평균 수명은 300일이라고 한다. 하지만 일부에서는 300일이 아니라는 의견도 있다. 그래서 해당 건전지 25개를 표본으로 조사하였더니, 310일의 평균수명이 나왔다. 어느 의견이 더 타당한지 유의수준 5%에서 검정하시오. 단 건전지의 모표준편차는 30으로 알려져 있다. 평균수명이 300일이 아니라는 의견이 나왔는데, 대소는 거론되지 않았으므로 대립가설은 "A사 건전지 1개의 평균 수명은 300일이 아니다"로 설정한다.

수식에 의해 검정통계량(Z값)을 구해보면 1.67이 나온다. 이 값이 임계치 내에 있는지 확인하기 위해 기각역을 구해보자.

$$= \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

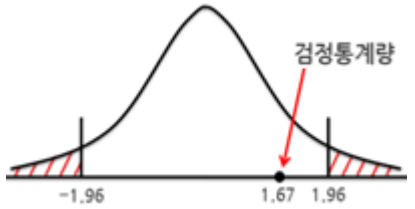
$$= \frac{310-300}{30/\sqrt{25}}$$

$$= 1.67$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.500000	0.503989	0.507978	0.511966	0.515953	0.519939	0.523922
0.1	0.539828	0.543795	0.547758	0.551717	0.555670	0.559618	0.563559
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568
0.3	0.617911	0.621720	0.625516	0.629300	0.633072	0.636831	0.640576
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802337	0.805105
0.9	0.815940	0.818589	0.821214	0.823814	0.826391	0.828944	0.831472
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976
1.2	0.884930	0.886861	0.888768	0.890651	0.892512	0.894350	0.896165
1.3	0.903200	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855
1.5	0.933193	0.934478	0.935745	0.936992	0.938220	0.939429	0.940620
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543
1.7	0.955435	0.956367	0.957284	0.958185	0.959070	0.959941	0.960796
1.8	0.964070	0.964852	0.965620	0.966375	0.967116	0.967843	0.968557
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002

0.975에 근접한 값

95% 신뢰수준에서 유의수준 α는 0.05인데, 양측검정이므로 α/2 = 0.025에 해당하는 값을 표준정규분포표에서 찾는다. 0.025는 정규분포의 오른쪽 면적에 해당한다. 하지만 정규분포표는 왼쪽 면적을 다루므로, 1 - 0.025 = 0.975에 해당하는 값을 정규분포표에서 찾아야 한다. 확률 0.975에 가장 가까운 Z값은 1.96인데, 양쪽으로 설정해야 하므로 기각역은 ±1.96이라는 것을 알 수 있다.



결론적으로, 검정통계량이 채택역 안에 위치하므로 "건전지의 평균수명은 300일이라고 할 수 있다." 라는 귀무가설이 채택된다.

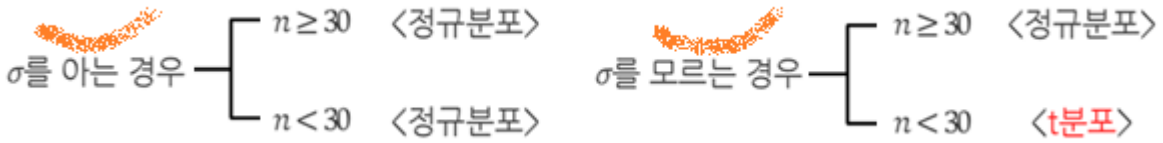
모평균의 가설검정 개념정리(σ를 모르는 경우)을 해보자.

σ를 모르는 경우에 대해서 알아보자. 보통 뭔가를 조사할 때, 모표준편차인 σ를 아는 경우는 흔치 않다. 그래서 모수를 모르기에 표본의 통계량을 사용하는데, σ를 모르는 경우에는 표본에서 얻어낸 표본표준편차 s를 사용한다.

σ를 아는 경우에는 정규분포를 사용하고, σ를 모르는 경우에는 t분포를 사용한다. 이때에 t분포로 확률을 구하는 것이 아니라, 단지 그래프의 x축 좌표를 구하는 데 활용할 뿐이다. T분포로는 기각역과 검정통계량을 구할 때 사용한다.

t분포는 표본의 수가 n≥30이면, 중심극한정리에 의해 정규분포와 값이 비슷해지기에 t분포 대신 정규분포를 사용할 수가 있다. n≥30이면 정규분포나 t분포나 값은 서로 비슷하기에, 꼭 정규분포를 사용해야 되는 것은 아니다. 하지만 표본의 개수가 30개 이상이면 정규분포를 사용하는 것이 일반적이다.

정규분포를 사용할 때 검정통계량 공식이 약간 다르다. 아무래도 모표준편차인 σ를 모르는 상태이기 때문에, 공식의 σ가 표본표준편차 s로 대체된다.



$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

<t분포>

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

<정규분포>

표본 31개 이하의 값 위주로 구성

v \ α	0,4	0,25
1	0.325	1.000
2	0.289	0.816
3	0.277	0.765
4	0.271	0.741
5	0.267	0.727
6	0.265	0.718
7	0.263	0.711
8	0.262	0.706
9	0.261	0.703
10	0.260	0.700
11	0.260	0.697
12	0.259	0.695
13	0.259	0.694
14	0.258	0.692
15	0.258	0.691
16	0.258	0.690
17	0.257	0.689
18	0.257	0.688
19	0.257	0.688
20	0.257	0.687
21	0.257	0.686
22	0.256	0.686
23	0.256	0.685
24	0.256	0.685
25	0.256	0.684
26	0.256	0.684
27	0.256	0.684
28	0.256	0.683
29	0.256	0.683
30	0.256	0.683
40	0.255	0.681
60	0.254	0.679
120	0.254	0.677
∞	0.253	0.674

〈t분포표〉

카이제곱검정 중 일원카이제곱

: 관찰도수가 기대도수와 일치하는 지를 검정하는 방법

: 종류 : 적합도/선호도 검정

- 범주형 변수가 한 가지로, 관찰도수가 기대도수에 일치하는지 검정한다.

적합도 검정

: 자연현상이나 각종 실험을 통해 관찰되는 도수들이 귀무가설 하의 분포(범주형 자료의 각 수준별 비율)에 얼마나 일치하는가에 대한 분석을 적합도 검정이라 한다.

: 관측값들이 어떤 이론적 분포를 따르고 있는지를 검정으로 한 개의 요인을 대상으로 함.

<적합도 검정실습>

주사위를 60 회 던져서 나온 관측도수 / 기대도수가 아래와 같이 나온 경우에 이 주사위는 적합한 주사위가 맞는가를 일원카이제곱 검정으로 분석하자.

주사위 눈금	1	2	3	4	5	6
관측도수	4	6	17	16	8	9
기대도수	10	10	10	10	10	10

<선호도 분석 실습>

5개의 스포츠 음료에 대한 선호도에 차이가 있는지 검정하기

이원카이제곱 - 교차분할표 이용

: 두 개 이상의 변인(집단 또는 범주)을 대상으로 검정을 수행한다.

분석대상의 집단 수에 의해서 독립성 검정과 동질성 검정으로 나뉜다.

독립성(관련성) 검정

- 동일 집단의 두 변인(학력수준과 대학진학 여부)을 대상으로 관련성이 있는가 없는가?
- 독립성 검정은 두 변수 사이의 연관성을 검정한다.

실습 : **교육수준과 흡연율 간의 관련성 분석 :** smoke.csv'

실습) 국가전체와 지역에 대한 인종 간 인원수로 독립성 검정 실습

두 집단(국가전체 - national, 특정지역 - la)의 인종 간 인원수의 분포가 관련이 있는가?

```
national = pd.DataFrame(["white"] * 100000 + ["hispanic"] * 60000 +
                        ["black"] * 50000 + ["asian"] * 15000 + ["other"] * 35000)
la = pd.DataFrame(["white"] * 600 + ["hispanic"] * 300 + ["black"] * 250 +
                        ["asian"] * 75 + ["other"] * 150)
```

cdf() : 누적분포함수

pdf() : 확률밀도함수

pmf() : 확률질량함수

이원카이제곱

동질성 검정 – 두 집단의 분포가 동일한가? 다른 분포인가? 를 검증하는 방법이다. 두 집단 이상에서 각 범주(집단) 간의 비율이 서로 동일한가를 검정하게 된다. 두 개 이상의 범주형 자료가 동일한 분포를 갖는 모집단에서 추출된 것인지 검정하는 방법이다.

동질성 검정실습1) 교육방법에 따른 교육생들의 만족도 분석 - 동질성 검정 survey_method.csv

동질성 검정 실습2) 연령대별 sns 이용률의 동질성 검정

20대에서 40대까지 연령대별로 서로 조금씩 그 특성이 다른 SNS 서비스들에 대해 이용 현황을 조사한 자료를 바탕으로 연령대별로 홍보 전략을 세우고자 한다.

연령대별로 이용 현황이 서로 동일한지 검정해 보도록 하자.

구분	F 사	T 사	K 사	C 사	기타	합
20대	207	117	111	81	16	532
30대	107	104	236	109	15	571
40대	78	76	133	32	17	336
합	392	297	480	222	48	1,439

독립성 검정은 두 변수 사이의 연관성을 검정하는데 비해,동질성 검정은 하위 모집단 사이 특정 변수에 대한 분포의 동질성을 검정한다.

집단 간 차이분석: 평균 또는 비율 차이를 분석

: 모집단에서 추출한 표본정보를 이용하여 모집단의 다양한 특성을 과학적으로 추론할 수 있다.

* T-test와 ANOVA의 차이

- 두 집단 이하의 변수에 대한 평균차이를 검정할 경우 T-test를 사용하여 검정통계량 T값을 구해 가설검정을 한다.
- 세 집단 이상의 변수에 대한 평균차이를 검정할 경우에는 ANOVA를 이용하여 검정통계량 F값을 구해 가설검정을 한다.

* 단일 모집단의 평균에 대한 가설검정(one samples t-test)

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$ <t분포>

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$ <정규분포>

실습 예제 1)

A중학교 1학년 1반 학생들의 시험결과가 담긴 파일을 읽어 처리 (국어 점수 평균검정) student.csv

실습 예제 2)

여아 신생아 몸무게의 평균 검정 수행 babyboom.csv

여아 신생아의 몸무게는 평균이 2800(g)으로 알려져 왔으나 이보다 더 크다는 주장이 나왔다.

표본으로 여아 18명을 뽑아 체중을 측정하였다고 할 때 새로운 주장이 맞는지 검정해 보자.

두 집단의 가설검정 - 실습 시 분산을 알지 못하는 것으로 한정하겠다.

* 서로 독립인 두 집단의 평균 차이 검정(independent samples t-test)

남녀의 성적, A반과 B반의 키, 경기도와 충청도의 소득 따위의 서로 독립인 두 집단에서 얻은 표본을 독립표본(two sample)이라고 한다.

실습) 남녀 두 집단 간 파이썬 시험의 평균 차이 검정

Male = [75, 85, 100, 72.5, 86.5]

female = [63.2, 76, 52, 100, 70]

실습) 두 가지 교육방법에 따른 평균시험 점수에 대한 검정 수행 two_sample.csv'

* 서로 대응인 두 집단의 평균 차이 검정(paired samples t-test)

처리 이전과 처리 이후를 각각의 모집단으로 판단하여, 동일한 관찰 대상으로부터 처리 이전과 처리 이후를 1:1로 대응시킨 두 집단으로 부터의 표본을 대응표본(paired sample)이라고 한다.

대응인 두 집단의 평균 비교는 동일한 관찰 대상으로부터 처리 이전의 관찰과 이후의 관찰을 비교하여 영향을 미친 정도를 밝히는데 주로 사용하고 있다. 집단 간 비교가 아니므로 등분산 검정을 할 필요가 없다.

실습) 복부 수술 전 9명의 몸무게와 복부 수술 후 몸무게 변화

baseline = [67.2, 67.4, 71.5, 77.6, 86.0, 89.1, 59.5, 81.9, 105.5]

follow_up = [62.4, 64.6, 70.4, 62.6, 80.1, 73.2, 58.2, 71.0, 101.0]

추론통계 분석 중 비율검정

- 비율검정 특징

  : 집단의 비율이 어떤 특정한 값과 같은지를 검증.

  : 비율 차이 검정 통계량을 바탕으로 귀무가설의 기각여부를 결정.

one-sample

A회사에는 100명 중에 45명이 흡연을 한다. 국가 통계를 보니 국민 흡연율은 35%라고 한다.

비율이 같냐?

two-sample

A회사 사람들 300명 중 100명이 커피를 마시고, B회사 사람들 400명 중 170명이 커피를 마셨다.

비율이 같냐?

세 개 이상의 모집단에 대한 가설검정 – 분산분석

'분산분석'이라는 용어는 분산이 발생한 과정을 분석하여 요인에 의한 분산과 요인을 통해 나누어진 각 집단 내의 분산으로 나누고 요인에 의한 분산이 의미 있는 크기를 크기를 가지는지를 검정하는 것을 의미한다.

세 집단 이상의 평균비교에서는 독립인 두 집단의 평균 비교를 반복하여 실시할 경우에 제1종 오류가 증가하게 되어 문제가 발생한다.

이를 해결하기 위해 Fisher가 개발한 분산분석(ANOVA, ANalysis Of Variance)을 이용하게 된다.

*** 서로 독립인 세 집단의 평균 차이 검정**

실습) 세 가지 교육방법을 적용하여 1개월 동안 교육받은 교육생 80명을 대상으로 실기시험을 실시. three_sample.csv'

사후검정(Post Hoc Test) : 연구가설이 채택되었다면...

구체적으로 어떤 교육방법 간에 차이가 있는지를 확인할 수 있다. (R의 경우 Tukey HSD() 등)

분산분석의 사후검정은 예를 들어 분산분석에서 '3가지 교육방법에 따른 실기시험의 평균에 차이가 있다.' 라는 결론이 나왔다면 구체적으로 어떤 교육방법 간에 차이가 있는지를 보여준다.

미국 수학자 John Tukey의 이름을 따서 명명 된 Tukey의 범위 테스트는 일원 분산 분석 (one-way ANOVA) 후에 사후 분석으로 사용되는 일반적인 방법이다. 이 테스트는 가능한 모든 쌍을 비교하며 예상되는 표준 오류보다 큰 두 가지 방법의 차이를 정확하게 식별하는 데 사용할 수 있다.