<통계 기본 지식>

**확률분포의 이해**

-어떤 변수 X의 값이 P(X)의 확률로 나오는 경우, X를 확률변수라고 한다.

-확률변수는 형태에 따라 이산확률변수와 연속확률변수가 있다.

이산확률변수는 정수와 같이 연속되지 않고 뚝뚝 끊어지는 값을 갖는 변수,

연속확률변수는 실수와 같이 연속되어 있는 값을 갖는 변수이다.

-확률분포는 확률변수가 특정 값을 가질 확률을 나타내는 함수로 표현한 것이다.

**연속확률분포**

-이산확률변수와 달리 연속확률변수는 연속된 값을 가지며 연속확률분포를 따른다.

-특정 구간 안에서 연속확률변수 X의 확률 P(X)는 연속확률분포 f(x)를 적분한 값과 같다.

(확률밀도곡선 : 68.2% 95.4% 99.7%)

**정규분포**

-정규분포는 우리 주변에서 쉽게 발견할 수 있는 연속확률분포 중 하나이다.

-정규분포는 모집단의 평균과 표준편차(분산의 양의 제곱근)를 알면 분포의 모양이 정해진다.

-평균이 같아도 표준편차가 다르면 분포의 모양을 서로 다르다.

-표준편차가 작을수록 평균 주변에 상대적으로 많은 데이터가 모이므로 뾰족한 종모양으로 그려지고, 표준편차가 넓게 퍼진 종모양으로 그려진다.

-이렇게 모두 다 다른 모양을 표준화시켜높은 것이 표준정규분포이다.

**표준정규분포**

-2개 이상의 확률변수가 정규분포를 따른다고 가정할 때, 두 확률변수는 평균과 표준편차가 서로 다르므로 현재 상태로 직접 비교하는 것은 쉽지 않다. 하지만 두 확률변수를 평균이 0이고, 표준편차가 1인 표준정규분포를 따르도록 표준화하면 직접 비교가 가능해진다.

-표준화 방법은 각 확률변수의 값에서 평균을 뺀 다음 표준편차로 나눈 z-score로 변환하는 것이다.

-z-score는 평균이 0, 표준편차가 1인 표준정규분포를 따른다.

**Chi-square분포**

-카이제곱 분포는 k개의 서로 독립적인 표준정규분포를 따르는 확률변수를 각각 제곱하여 모두 더한 값으로 얻은 분포이다.(k는 카이제곱 분포의 자유도)

-자유도에 따라 분포의 모양이 크게 달라지는데, 자유도가 증가할수록 정규분포에 가까워진다.

스튜던트 t분포

-스튜던트 t분포는 정규분포하는 표본의 평균으로 모집단의 평균을 추정할 때 사용되는 분포이다.

-t분포는 자유도의 크기가 커질수록 모양이 표준정규분포에 가까워진다.

-t검정 및 회귀계수의 유의성 검정을 할 때 t통계량이 사용된다.

-sample수가 증가할수록 정규분포에 가까워진다.

F분포

-F분포는 카이제곱 분포를 따르는 두 개의 확률값을 각각의 자유도로 나눈 후 다시 나눈 값으로 얻은 분포이다.

-F분포는 ANOVA에서 집단 간 변동과 집단 내 변동의 평균을 비교할 때 활용된다.

-아울러 회귀모형의 유의성 검정을 할 때 F분포가 활용된다.

**통계적 가설검정**

-통계적 가설검정은 어떤 주장(가설)에 대해 모수를 이용하여 해당 주장이 합당한 것인지를 판단하는 과정이다.

(모수는 모집단의 특성인데, 전체 모집단 대신 표본을 추출하여 표준의 특성으로 모수를 추정한다.)

-통계적 가설검정은 다음과 같은 단계를 거친다.

>귀무가설과 대립가설을 설정

>유의수준과 기각역을 설정(유의수준은 참인 귀무가설을 기각하는 위험)

>검정통계량 및 유의확률을 계산(표본의 확률분포에 따라 유의확률이 결정)

>유의확률이 기각역 바깥에 위치하면(즉 기각역보다 작으면)귀무가설을 기각

**상관성 분석**

컴퓨터의 성능은 크게 향상되었고, 빅데이터 시대가 도래함에 따라 많은 기업들이 자산으로서 데이터의 중요성에 대해 인식하게 되었고, 그 결과 기업은 과거보다 많은 데이터 규모를 확보하게 되었다.

-그 결과 과거에 비해 모형 적합에 사용할 수 있는 입력변수(input variable)의 수가 크게 증가하였지만, 이는 연산 비용의 증가로 연결된다.

-적합도가 높은 모형의 생성을 위해, 목표변수(target variable)와 상관성이 있는 입력변수를 선택함으로써 작업 효율을 높일 수 있다.

상관성 분석의 종류(불필요한 연산 비용을 빼기위해 필수!)

-입력변수와 목표변수의 종류에 따라 4가지로 구분할 수 있다.

목표변수가 연속형이면 회귀모형, 범주형이면 분류모형

>목표변수:연속형, 입력변수:연속형 🡪 상관계수

>목표변수:연속형, 입력변수:범주형 🡪 t-test/ANOVA (차이나야 상관성)

>목표변수:범주형, 입력변수:범주형 🡪 카이제곱

**상관계수**(목표변수와 입력변수 모두 연속형)

-두 변수가 모두 연속형일 때 상관계수를 통해 상관관계를 확인할 수 있다.

-상관계수(p, rho)는 두 변수 간 선형관계가 있는지 정도를 나타내는 지표

>상관계수는 -1 <= p <= 1

>상관계수가 -1 또는 1에 가까울수록, 두 변수는 선형관계가 크다.

>상관계수가 0에 가까울수록, 두 변수는 선형관계가 없다.

>하지만 두 변수 간 직선의 관계가 있는 판단하는 기준은 상관계수 그 자체가 아니라, 상관분석을 실행한 결과로 생성되는 유의확률이다.

(유의확률이 0.05보다 작을 때 두 변수간 선형관계가 있다고 판단)

\*\*두 변수간 ‘상관관계’가 있다고 해도 그것이 ‘인과관계’를 의미하는 것은 아니다!!

**상관계수의 종류**

-피어슨 상관계수 : 두 연속형 변수간 선형관계 측정

두 연속형 변수의 공분산을 각각의 표준편차로 나누어

표준화한 것

상관계수가 1이면 양의 직선관계, -1이면 음의 직선관계

ex)국어점수가 높은 학생이 영어점수도 높다?

Cor()함수

-스피어만 상관분석 : 이산형(정수) 또는 순서형인 두 변수 간 관계 측정

한 변수가 증가할 때 다른 변수도 증가하는지 확인

연속형 데이터가 주어지면 순서형으로 변환하여

상관계수를 계산한다.

스피어만 상관계수의 공식은 피어슨 상관계수의 공식

과 비슷하다.

스피어만 상관계수가 1이면 한 변수가 증가할 때 다른

변수도 반드시 증가한다는 것

ex)국어시험 등수와 영어시험 등수(‘순서’가 같은지)

Cor( , method = ‘spearman’)함수

이 때 유의확률까지 확인하려면 cor.test() 함수

p-value가 0.05보다 작으면 귀무가설 기각 🡪 상관관계 있다.

\*\*대체로 피어슨이 높으면 스피어만도 높다.

-정준상관분석(CCA, canonical correlation analysis)

2개 이상의 변수들로 구성된 2개의 집단 간 상관계수를 측정한 것

상관계수 측정을 위해 각 집단을 대표하는 주성분(principal compenent)을 생성한다. (주성분은 각 집단을 구성하는 변수들의 선형관계를 이룬다.)

2개의 집단 간 상관계수는 제 1 주성분으로 측정했을 때 가장 큰 값을 가진다.

Ex)집단A(월소득, 부양가족 수)와 집단 B(보유카드 수, 카드 이용금액)의 관계

\*\*집단A의 제 1주성분, 집단B의 제 1주성분을 뽑아서 비교하는 것

\*\*\*주성분 분석: 여러 차원의 컬럼들을 그룹화하기 위해, 각 그룹마다 크게 고려하는 종목을 다르게 해서 비교하는 것

**t-test**

목표변수는 연속형, 입력변수는 명목형(2개의 level)일 때

-독립 2표본 t-검정은 2개 집단의 평균이 서로 같은지 여부를 판단할 때 사용된다.

- 독립 2표본 t-검정은 두 집단 내 변화량을 고려하여 두 집단 간 평균이 통계적으로 유의미한 차이를 가지는지 검증하는 방법이다.

>두 집단간 평균의 차이가 작을수록, 두 집단은 서로 유사하다.

>두 집단간 평균의 차이에 대한 표준오차가 작을수록, 두 집단의 평균을 객관적으로 비교할 수 있다.

**t-test의 가정**

-독립 2표본 t-검정은 실행하기에 앞서 데이터가 독립성, 정규성 및 등분산성 등 3가지 조건을 만족하는지 여부를 확인해야한다.

(입력변수의 레벨이 서로 독립이어야하 하며, 목표변수는 정규분포하고 레벨 간 분산이 같아야 한다.)

-독립성 조건은 데이터가 수집되는 과정에서 두 집단 간 서로 아무런 관계가 없다는 것을 분석가가 판단할 수 있다.

>정규성 조건 Shapiro.test()함수를 실행하여 검정

-귀무가설은 ‘데이터가 정규분포한다.’이므로 유의확률이 0.05보다 커야함

-정규성 만족 못하면, ‘Wilcoxon 순위합 검정’(비모수적)

>등분산성 검정 var.test()함수를 사용

-귀무가설은 ‘두 집단의 분산이 같다’이므로 유의확률이 0.05보다 커야함

-등분산성 만족 못하면, ‘Welch’s t-test’으로 실행해야 하는데, t.test()함수의 ‘var.equal’인자에 FALSE를 할당하면 ‘Welch’s t-test’가 실행된다.(이분산)

>t-검정은 t.test()함수 사용

**ANOVA(분산분석)**

-분산분석은 3개 이상의 집단 간 평균이 다른지 확인할 때 사용된다.

-목표변수 1개, 입력변수의 level이 3개 이상일 때 일원분산분석을 사용할 수 있다.

-분산분석은 편차제곱합을 각각의 자유도로 나누어서 얻은 평균제곱합의 비를 검정 통계량으로 사용한다.

(유의확률이 0.05보다 작으면 ‘모든 집단의 평균이 같다.’는 귀무가설 기각)

-분산분석은 독립성, 정규성, 등분산성 가정을 만족해야한다.

-정규성 가정은 Shapiro.test()함수 이용

(정규성 가정을 만족하지 못하면 ‘Kruskal-Wallis 순위합 검정’

-입력변수의 레벨이 3개 이상인 등분산성 검정은 ‘Levene검정’ 실행

(‘각 처리별 분산이 같다’는 귀무가설, 0.05보다 크면 만족)

-등분산 검정은 입력변수의 레벨별 목표변수가 등분산인지 확인하는 것이다.

(‘집단 간 분산이 같다’는 귀무가설, 0.05보다 커야 등분산 가정 만족)

levene.test()사용

-분산분석은 oneway.test()함수 사용

(유의확률이 0.05보다 작아야 ‘모든 집단의 평균이 같다’는 귀무가설을 기각할 수 있다.)

**카이제곱 검정**

입력변수와 목표변수가 모두 범주형일 때

-카이제곱 검정은 입력변수의 범주에 따라 목표변수의 빈도수 차이가 통계적으로 유의한지 검정할 때 사용된다.

-입력변수 별로 목표변수의 실제값-기대값의 차이가 클수록 카이제곱 통계량도 커딘다.

(카이제곱 통계량이 증가하면 유의확률이 감소)

-crosstable()함수 사용

-귀무가설이 ‘집단 간 빈도수의 차이가 없다’이므로 유의확률이 0.05보다 작아야 차이가 있는 것이다.