**교차분석 & 상관분석**

**교차분석(CrossValidation)**

**Theory**

교차분석이란 명목척도 혹은 순서척도와 같은 **범주형 자료**들 간의 상호 연관성을 알아볼 때 사용하는 방법이다. 두 범주형 변수에 대한 관련성을 파악하고 이를 통계적으로 검정하여 복잡한 상황에 대한 통찰력을 얻을 수 있다. (교차분석이란 적합성, 독립성, 동질성을 검정하는 작업)

보통 **적합성 검정, 독립성 검정, 동질성 검정**에 사용되며, **카이제곱(x^2) 검정 통계량**을 이용한다.

교차분석에서는 두 변수의 각 범주를 교차하여 데이터의 빈도수를 표로 정리한 교차표를 사용하는데, 교차분석은 이러한 교차표를 통해 각 셀의 교차빈도와 기대빈도간의 차이를 검정할 때 사용된다.

* 목표 : n개의 표본을 k개의 범주로 분리한 뒤, 각 범주의 관측도수(O)와 주어진 확률 분포에 의거한 기대도수(E)들이 적합한지 여부를 검정.
* 귀무가설 : 실제 분포와 이론적 분포 간의 차이가 없다.
* 대립가설 : 실제 분포와 이론적 분포 간의 유의미한 차이가 있다.

**적합성 검정 ( 범주형 변수 1개로 하는 작업 ) : 한 개의 특성에 대해 특정 분포를 따를지 검정**

적합성검정은 변주형인 하나의 변수에 대해, 이 변수가 우리가 기대하는 어떤 분포를 따르는지 여부를 검정하는 작업. 실제로 관측된 값과 일어날 것으로 기대하고 있는 값을 비교하는 검정.

* 카이제곱 검정 : 카이제곱 적합도 검정은 변수가 지정된 분포를 따를 가능성이 있는지 여부를 확인하는데 사용되는 통계적 가설 검정입니다. 귀무가설은 두 변수가 지정된 분포를 따른다이며, 카이제곱 통계량이 클수록 적합도가 낮으며 유의확률이 감소합니다. (df=k-1 | k=범주수)

**독립성 검정 ( 두 개의 범주형 변수로 하는 작업 ) : 두 변수 간의 연관성, 독립성 검증**

독립성 검정은 모집단이 두 개의 변수 A,B에 의해 범주화되었을 때, 이 두 변수들 사이의 관계가 독립인지 아닌지를 검정하는 방법. 독립이란 A,B 사건이 있을 때, 한 사건이 일어날 확률이 다른 사건에 영향을 주지 않는 상태를 의미함.

* 카이제곱 검정 : 귀무가설은 두 변수 간의 독립성이 존재한다. 연관이 없다.

**동질성 검정 ( 두 개의 범주형 변수로 하는 작업 ) : 서로 다른 모집단에서 추출.**

동질성 검정은 모집단이 임의의 변수에 따라 R개의 속성으로 범주화되었을 때, R개의 부분 모집단에서 추출한 각 표본이 C개의 범주로 또 다시 범주화될 때 집단의 분포가 동일한지 검증하는 방법.

본 검증은 위 독립성 검증과 같은 방법으로 진행됨.

**상관분석(Correlation)**

**Theory**

상관분석이란 두 변수 간 관계의 정도를 알아보기 위한 분석방법. **(인과관계 X | 독립성 보장 X)**

상관계수는 데이터 간 상관관계가 얼마나 강한지를 수치화한 값으로 두 변수 간 관련성의 정도를 의미하며, 일반적으로 피어슨(Pearson), 스피어만(Spearman) 상관계수를 많이 사용합니다.

공분산은 두 확률변수가 함께 변화하는지의 정도를 측정하는 값으로 상관성의 경향까지는 파악할 수 있지만, 측정 단위에 따라 값이 달라지므로 절대적인 상관성의 정도를 파악하기에는 한계가 있다. 따라서 공분산을 두 변수의 표준편차 곱으로 나누어 표준화한 상관계수를 이용해 상관성을 파악한다.

상관계수는 양과 음이 존재하며 **0.3**과 **0.7**을 기준으로 약하고 강한 상관관계가 파악된다.

**Correlation Method**

* 피어슨 상관계수(Pearson) : **연속형, 정규분포**

두 연속형 자료가 모두 정규성을 따른다는 가정 하에 선형적 상관관계를 구하는 방법.

* 스피어만 상관계수(Spearman) : **비정상성 연속형, 범주형**

두 변수가 정규성을 만족하지 않거나, 순위 및 순서 형태로 주어지는 경우 스피어만 상관계수를 사용하여 데이터 간의 관계를 파악한다. 스피어만은 모수를 측정하는 대신 데이터에 순위를 매긴 후 그 순위에 대한 상관계수를 산출하는 비모수적 방법이다. 피어슨과 달리 비선형 관계의 연관성까지도 파악할 수 있시 때문에 이산형, 순서형 변수에도 사용이 가능하다.

* 켄달 상관계수(Kendall) :

켄달 상관계수는 데이터가 (X,Y) 순서쌍으로 주어졌을 때 X가 커짐에 따라 Y도 커지는 부합과 반대로 Y가 작아지는 비부합 간의 비율로 상관계수를 산출한다.

**Correlation Test**

* 귀무가설 : 두 변수 간의 상관관계는 존재하지 않는다.
* 대립가설 : 두 변수 간의 상관관계가 존재하며, 얼마나 어떻게 존재한다.