**T-test & Data Sampling**

**Data Sampling**

**Theory**

표본조사에서 중요한 점은 모집단을 대표할 수 있는 표본의 추출이며, 표본 추출 방법에 따라 분석 결과에 큰 차이가 발생한다.

**Sampling Strategy**

* 단순 임의 추출법(Simple random sampling): 단순 임의 추출은 전체 데이터에서 데이터를 선택할 확률을 모두 동일하게 설정하여 표본을 추출하는 방법입니다. 일반적으로 train\_test\_split시 가장 많이 쓰이는 표본추출방법입니다.
* 층화 추출법(Strarified random sampling): 모집단이 이질적인 몇 개의 계층으로 이루어져 있을 때 모든 계층으로부터 원소를 임의로 추출하여 각 계층을 고루 대표할 수 있도록 랜덤하게 표본을 추출하는 방법입니다.

**T-검정(T-test)**

**Theory**

T-검정은 두 집단의 평균을 통계적으로 비교하기 위해 사용되는 검정방법이다. 어떤 방식으로 집단의 평균을 비교하느냐에 따라 일표본 t-검정, 대응표본 t-검정, 독립표본 t-검정으로 나누어진다. 일반적으로 가설검정에 의해 산출된 유의확률(p-value)이 유의수준(0.05)보다 작을 경우 귀무가설을 기각하고, 대립가설을 채택한다. 반대로 클 경우, 대립가설을 기각한다.

* 귀무가설 : 현재까지 일반적으로 주장되어 온 주장
* 대립가설 : 데이터 표본으로부터 확실한 증거를 바탕으로 증명하고자 하는 가설



단일표본, 쌍으로 된 두 표본은 정규성 검정을 실시하고 만족할 경우 t-test를 진행, 불만족할 경우, Wilcoxon test를 진행.

독립적인 두 표본은 정규성 검정과 등분산성 검정을 모두 실시함. 그리고 **위 케이스 모두 이전에 이상치 처리를 해주어야함.**

**정상성 검정**

* 콜모고로프 스미르노프 검정(Kolmogorov Smirnov)

Kolmogorov Smirnov 검정은 표본 수가 많을 때 2개의 분포의 **동질성을 확인하는 검정**입니다. 귀무가설은 **두 분포가 동일하다(정규성을 따른다)**.

* 샤피로-윌크 검정(Shapiro-Wilk normality test)

Shapiro Wilk 검정은 표본 수가 많지 않을 때(50개 미만) 사용하는 **정상성 검증 방법**입니다. 귀무가설은 정규분포를 따른다.

**등분산성 검정**

* Bartlett, fligner, levene

두 독립분포의 등분산성을 검정하는 방법으로 귀무가설은 두 독립분포가 등분산성을 가진다.

**일표본 T-검정**

단일모집단에서 관심이 있는 연속형 변수(하나!)의 평균값을 특정 기준값과 비교하고자 할 때 사용돼는 검정방법으로, **모집단의 구성요소들이 정규분포를 이룬다는 가정(정상성) 하에** 검정통계량(p-value)을 계산한다.

* 귀무가설 : 모평균의 값은 설정한 기준값이다.
* 대립가설 : 모평균의 값은 설정한 기준값과 다르다.

▷ 전략! : 이상치 처리 -> 정상성 검증(stats.shapiro) -> ttest\_1samp or wilcoxon 검증 시행

**대응표본 T-검정**

단일모집단에 대해 두 번의 처리(두 번의 평균값 추정)을 가했을 때, 두 결과에 대한 평균의 차이를 비교하고자 할 때 사용하는 검정방법이다. 하나의 모집단에서 크기가 n개인 하나의 표본을 추출한 후, 각기 다른 두 번의 측정을 실시한다. 즉, 모수가 두 개 추정된다. 각 개체별로 짝 지어진 관측값 사이에 차이가 있는지를 검증하는 검정방법. (u0 = u1)

* 모수 : 두 개의 모평균의 차이
* 귀무가설 : 두 개의 모평균의 차이는 없다.
* 대립가설 : 두 개의 모평균의 차이가 있다.

▷ 전략! : 이상치 처리 -> 정상성 검증(stats.shapiro) -> ttest\_rel or wilcoxon 검증 시행

**독립표본 T-검정**

두 개의 독립된 모집단의 평균을 비교하고자 할 때 사용하는 검정방법이다. **두 개의 모집단에서** 크기가 n개인 표본을 각각 추출한 후 표본의 관측값들을 이용해 검정을 실시한다. 따라서 독립표본 t-검정에서는 모집단, 모수, 표본이 모두 두 개씩 존재한다.

두 모집단은 정규성과 독립성 그리고 등분산성을 모두 만족해야한다. **특히, 등분산성 만족 여부에 따라 다른 계산법이 사용된다.**

* 모수 : 서로 독립된 두 모집단의 평균
* 귀무가설 : 서로 독립된 두 모집단의 평균은 같다.
* 대립가설 : 서로 독립된 두 모집단의 평균은 다르다.

▷ 전략! : 이상치 처리 -> 등분산성 검증(levene) -> ttest\_ind or mannwhitneyu 검증 시행