21.02.26 통계

노트북: [면접 관련]
만든 날짜: 2021-02-26 오전 10:17 **수정한 날짜:** 2021-03-03 오후 9:38
작성자: 황인범
URL: https://www.google.com/search?q=T+%EA%B2%80%EC%A0%95&oq=T+%EA%B...

---

1. 통계 지식
    1. 탐색적 데이터 분석
        1. 정형화된 데이터 요소
            1. 연속형 - 어떤값이든 들어가는 데이터(온도 같은거)
            2. 이산형 - 정수값 (사건의 발생 횟수)
            3. 범주형 - 범주안의 값만 가능
            4. 이진형 - 0 또는 1
            5. 순서형 - 순위가 있는 범주형 데이터
        2. 테이블 데이터
            1. 데이터 프레임 - 테이블형데이터 구조
            2. 피처 - 하나의 열
            3. 레코드 - 하나의 행
            4. 결과 - 결과
        3. 위치추정
            1. 평균 - 총합을 개수로 나눈 것
            2. 가중평균 - 가중치를 곱한 값의 총합
            3. 중간값 - 데이터에서 가장 가운데 위치 값
            4. 가중 중간값 - 가중치 값을 위에서부터 더할때 중간값
            5. 절사평균 - 정해진 개수의 극단값을 제외한 나머지 값들의 평균 (양 끝에 몇개 삭제)
            6. 로버스트하다 - 극단값에 민감하지 않다
            7. 특잇값 - 대부분의 값과 매우 다른 데이터 값
        4. 변이 추정
            1. 변이는 데이터 값이 얼마나 밀집해 있는지 혹은 퍼져 있는지를 나타내는 산포도
            2. 편차 - 관측값과 추정값의 차이
            3. 분산 - 평균과 편차를 제곱한 값들의 합을 n-1로 나눈 값
            4. 표준편차 - 분산의 제곱근
            5. 평균절대편차 - 평균과 편차의 절댓값의 평균
            6. 중간값의 중위절대편차 - 중간값과의 편차의 절대값의 중간값
            7. 범위 - 데이터의 최댓값과 최솟값의 차이
            8. 순서통계량 - 최소에서 최대까지 정렬된 데이터 값에 따른 계량형
            9. 백분위수
            10. 사분위범위 - 75-25 사이의 차이(IQR)
        5. 데이터 분포 탐색
            1. boxplot(상자그림)
            2. 도수분포표 - 엑셀같은거
            3. 히스토그램 - 작대기로 나와있는거
            4. 밀도그램 - 히스토그램 곡선으로 만든 것
        6. 이진 데이터와 범주 데이터 탐색
            1. 용어정리
                1. 최빈값 - 가장 많이 등장하는 값
                2. 기댓값 - 범주의 출현 확률에 따른 평균
                3. 막대도표 - 막대로 나타낸 그림
                4. 파이그림 - 부채꼴모양으로 나타낸 그림
        7. 상관관계
            1. 용어정리

1. 상관계수 - 수치적 변수들간에 어떤 관계가 있는지를 나타내기 위해 사용되는 측정량
2. 상관행렬 - 변수들의 상관관계를 나타내는 표
3. 산점도 - x축과 y축이 서로 다른 두개의 변수를 나타내는 도표

8. 두 개 이상의 변수 탐색하기
    1. 용어정리
        1. 분할표 - 두가지 이상의 범주형 변수의 빈도수를 기록한 표
        2. 유각형 구간 - 두 변수를 육각형 모양의 구간으로 나눈 그림
        3. 등고 도표 - 두 변수의 밀도를 등고선으로 표시한 도표
        4. 바이올린 도표 - 상자그림과 비슷하지만 밀도추정을 함께 보여준다.(그 이상한 타원같은거)

2. 데이터와 표본분포
    1. 랜덤표본추출과 표본편향
        1. 용어정리
            1. 표본 - 부분집합
            2. 모집단 - 데이터 집합을 구성하는 전체 대상 혹은 전체 집합
            3. N - 모집단의 크기
            4. 임의표집 - 무작위로 표본을 추출하는 것
            5. 층화표집 - 모집단을 층으로 나눈 뒤, 각 층에서 무작위로 표본을 추출하는 것
            6. 단순임의표본 - 모집단 층화 없이 랜덤표본추출로 얻은 표본
            7. 표본편향 - 모집단을 잘못 대표하는 표본
    2. 선택 편향
        1. 용어정리
            1. 편향 - 계통적 오차
            2. 데이터스누핑 - 광범위하게 데이터를 살피는 것
            3. 방대한 검색효과 - 중복 데이터 모델링이나 너무 많은 예측변수를 고려하는 모델링에서 비롯되는 편향 혹은 비현재성
    3. 통계학에서의 표본분포
        1. 표본통계량 - 더 큰 모집단에서 추출된 표본 데이터들로부터 얻은 측정 지표
        2. 데이터 분포 - 어떤 데이터 집합에서의 각 개별 값의 도수 분포
        3. 표본분포 - 여러 표본들 혹은 재표본들로부터 얻은 표본 통계량의 도수 분포
        4. 중심극한정리 - 표본크기가 커질수록 표본분포가 정규분포를 따르는 경향
        5. 표준오차 - 표본 통계량의 변량
    4. 부트스트랩
        1. 용어
            1. 부트스트랩 표본 - 관측 데이터 집합으로부터 얻은 복원추출 표본
            2. 재표집 - 관측 데이터로부터 반복해서 표본추출하는 과정, 부트스트랩과 순열 과정을 포함한다.
    5. 신뢰구간
        1. 용어정리
            1. 신뢰수준 - 신뢰구간의 백분율로서 관심 통계량을 포함할 것 으로 예상되는
            2. 구간 끝점 - 신뢰구간의 최상의 , 최하위 끝점
    6. 정규분포
        1. 용어
            1. 오차 - 데이터 포인트와 예측값 혹은 평균 사이의 차이
            2. 표준화하다 - 평균을 빼고 표준편차로 나눈다
            3. z 점수 - 개별 데이터 포인트를 정규화한 결과
            4. 표준정규분포 - 평균 0, 표준편차=1 인 정규분포

5. QQ그림 - 표본분포가 정규분포에 얼마나 가까운지를 보여주는 그림

7. 긴꼬리분포
    1. 용어
        1. 꼬리 - 적은 수의 극단값이 주로 존재하는 , 도수분포의 길고 좁은 부분
        2. 왜도 - 분포의 한쪽 꼬리가 반대쪽 다른 꼬리보다 긴 정도
8. 스튜던트의 t 분포
    1. 정규분포와 생김새가 비슷하지만, 꼬리 부분이 약간 더 두껍고 길다. 이것은
       표본 통계량의 분포를 설명하는데 광범위하게 사용
9. 이항분포
    1. 용어정리
        1. 시행 - 독립된 결과를 가져오는 하나의 사건
        2. 성공 - 시행에 대한 관심의 결과
        3. 이항식 - 두가지 결과 (ex 0/1 , 예/아니오)
        4. 이항시행 - 두 가지 결과를 가져오는 시행
        5. 이항분포 - x번 시행에서 성공한 횟수에 대한 분포
10. 푸아송분포와 그외 관련 분포
    1. 용어정리
        1. 람다 - 단위 시간이나 단위 면적당 사건이 발생하는 비율
        2. 푸아송 분포 - 표집된 단위 시간 혹은 단위 공간에서 발생한 사건의 도수 분포
        3. 지수 분포 - 한 사건에서 다음 사건까지의 시간이나 거리에 대한 도수분포
        4. 고장률 추정 - 드물게 발생하는 사건의 경우
        5. 베이불 분포 - 사건 발생률이 시간에 따라 변화하는 지수분포의 일반화된 버전

3. 통계적 실험과 유의성 검정
    1. A/B 검정
        1. 두 처리 방법 혹은 , 제품등 어느 쪽이 다른 쪽보다 우월하다는 것을 입증하기 위해 실험군을 두 그룹으로 나누어 진행하는 실험
        2. 용어정리
            1. 처리 - 어떤 대상에 주어지는 특별한 환경이나 조건
            2. 처리군 - 특정 처리에 노출된 대상들의 집단
            3. 대조군 - 어떤 처리도 하지않은 대상들의 집단
            4. 임의화 - 처리를 적용할 대상을 임의로 결정하는 과정
            5. 대상 - 처리를 적용할 개체 대상
            6. 검정 통계량 - 처리효과를 측정하기 위한 지표
        3. 그룹 A/B를 비교하는데 사용하는 검정통계량 또는 측정 지표에 주의를 기울여야함
    2. 가설검정
        1. 관찰된 효과가 우연에 의한 것인지를 알아내는 것 / 가설에 따른 신뢰구간을 어떻게 잡을지 구하고 유의확률을 구하고
        2. 가설을 세우고 검정통계량을 구해 가설이 맞는지 아닌지를 판단하는 검정방법입니다.
        3. 단계
            1. 귀무가설 / 대립가설 설정 및 유의수준 결정
            2. 검정통계량 결정(어떤 검정방법을 쓸건지 결정)
            3. 기각역 결정
            4. 검정통계량의 계산
            5. 통계적 의사 결정
        4. 용어정리
            1. 귀무가설 - 우연 때문이라는 가설 (기존의 가설)
                1. 귀무가설이 틀렸다는 걸 증명해야 내 대립가설이 인정된다.
            2. 대립가설 - 귀무가설과의 대조되는 가설
                1. 귀무 가설 a>b => 대립가설 a <= b
            3. 일원검정 - 한방향으로만 우연히 일어날 확률을 계산하는 가설검정

4. 이원검정 - 양방향으로 우연히 일어날 확률을 계산하는 가설 검정

3. 재표본 추출
    1. 표본을 반복적으로 추출하는 것
    2. 용어정리
        1. 순열검정 - 두 개 이상의 표본을 함께 결합하여 관측값들을 무작위로 재표본으로 추출하는 과정
            1. 관찰된 차이가 순열로 보이는 차이의 집합밖에 있다면 이것은 통계적으로 의미가 있다고 한다.
        2. 복원/비복원 - 다시 넣을지 말지
    3. 통계적 유의성과 p 값(유의확률)
        1. 기각여부를 판단하는 확률 값
        2. 귀무가설이 틀렸다는 결과가 관측될 확률
        3. p-value 는 유의확률 / 95퍼센트는 알파이고 유의수준이다.
        4. 통계적유의성이란 실험 결과가 우연히 일어난 것인지 아니면 우연히 일어날 수 없는 극단적인 것 인지를 판단하는 방법
        5. 우연히 벌어질 수 있는 변동성의 바깥에 존재한다면 우리는 이것을 통계적으로 유의하다고 말한다.
        6. 용어정리
            1. p-value - 귀무가설을 구체화한 기회모델이 주어졌을때 , 관측된 결과와 같이 특이하거나 극단적인 결과를 얻을 확률
            2. 알파 - 실제 결과가 통계적으로 의미있는것으로 간주되기 위해 우연에 의한 기회 결과가 능가해야 하는 비정상적인 가능성의 임계확률
            3. 제 1종오류 - 귀무가설이 맞는데 틀렸다고 할 확률
            4. 제 2종 오류 - 귀무가설이 틀렸는데 맞다고 할 확률
        7. 유의수준 - 5%, 1%를 많이 사용한다. - 극단적인 확률이 나올 확률
    4. t 검정
        1. t-분포를 따르는 통계적 가설 검정법 / 두 집간의 평균을 비교하는 모수적 검정
        2. 용어정리
            1. 검정 통계량 - 관심의 차이 또는 효과에 대한 측정 지표
            2. t 통계량 - 표준화된 형태의 검정 통계량
            3. t 분포 - 관측된 t 통계량을 비교할 수 있는 기준분포
    5. 카이제곱 분포 - 모분산에 대한 가설 검정
    6. F분포 - 분산의 동일성에 대한 검증
4. 회귀와 예측
5. 분류
6. 통계적 머신러닝
7. 비지도 학습
8. 이러한 검정법들은 내가 어떤 가설을 세우냐에 따라 다르게 쓰여지는 걸로 알고 있습니다.

1종오류 - 귀무가설을 잘못 기각하는 확률 (알파 오류)
2종오류 - 귀무가설을 잘못 채택하는 확률(베타 오류)
검출력(r) = 1- (제 2 종 오류의 확률)b