데이터 분석과 관련된 통계학 주요내용 복습

📌 통계와 관련된 비전문가나 주요 내용 복습이 필요한 사람에게 데이터의 통계적 해석에서 겪는 혼란을 주는 주요 통계 개념을 설명하기 위해 내용을 선별하고 공부함

▼ **연속형 변수와 불연속형 변수에 대한 이해**

- 특정한 카테고리로 묶일 수 있으면 (의미를 담고 있으면) **불연속변수** (Discrete Variable)
- 어떤 값이든 될 수 있는 것이면 **연속변수** (Continuous Variable)

▼ **분포 (Distribution) 에 대한 이해**

> 발생하는 여러 결과에 대한 확률을 제공하는 수학적 함수

> 대수의 법칙에 의해 데이터는 '분포'에 가까워질 수 있음

> 데이터가 설명하는 '분포'를 알아내기 위한 목적은, 함수를 알면 그 현상을 설명가능하고, 예측가능하기 때문임

- 차트나 종모양 곡선을 흔히 상상하지만 그것은 '분포'가 아님을 기억하자.
- 불연속변수에 대한 자료는 히스토그램으로, 연속형변수에 대한 자료는 흔히 라인 차트로 확인하게 됨
- **연관학습 - 대수의 법칙**
    - 측정 대상의 숫자 또는 측정 횟수가 많아질수록 실제의 결과가 예상된 결과에 가까워진다는 경험적 확률과 수학적 확률 사이의 관계를 나타내는 법칙
    - 표본집단의 크기가 커지면 그 표본평균이 모평균에 가까워짐을 의미하므로, 표본 수가 많을수록 통계적 정확도가 올라가게 됨
- 즉, 데이터가 더 정확하고 표본집단의 크기가 커질수록 '분포'라는 현상을 더 정확하게 근사할 수 있음
- 따라서, 실제 결과가 되는 데이터를 통해 '분포'라고 하는 함수를 추정하는 것이 통계학에서 분포를 보고자 하는 주요 이유
- 이는, 함수를 알면 그 현상을 설명가능하고, 재현가능하기 때문임
- 불

▼ **표준편차 (Standard Deviation) 에 대한 이해**

> '데이터가 얼마나 흩어져있는가?'를 알기 위한 지표로, 분산=(편차제곱합/표본수)의 제곱근을 취한 값

> 데이터의 중심이나 중심경향성을 알기위한 대푯값인 '평균'에서 얼마나 떨어져있는지를 알기 위한 것

> 표준편차라고 하는 '분산에 제곱근을 취한 값'을 별도로 쓰는 이유는, 분산에서 제곱을 통해 측정 단위가 없어진 부분을 다시 복원하여 해석할 수 있기 때문

- **분산이란?**
    - 데이터가 평균으로부터 거리가 멀수록 값이 커지고, 거리가 가까울수록 값이 작아지는 지표
    - 다양한 곳에서 분산이 사용되나, 기초적인 기술통계에서는 자료 해석의 용이함을 위해 표준편차를 주로 사용함
    - **분산 = 편차제곱합/표본수 로 계산되는 이유**
        - 평균으로부터 떨어진 거리를 단순히 평균과 측정값 간의 차이로 계산하여 표본수로 나누면 항상 0의 값을 가지게 됨
        - 이를 해결하기 위해 각 거리마다 절대값과 같이 양수로 계산되도록 해야 분산을 알 수 있음
        - 모듈이나 다른 방법을 통해 각 편차를 양수로 만들어주는 방법이 있으나, 미분/적분 가능한 상태를 유지하기 위해 (절댓값 그래프처럼 뾰족한 경우 뾰족점에서 미분불가능한 것과 같이) 제곱하는 경우가 가장 적절함
        - 즉, 모든 값이 양수가 되도록하면서 양수로 변환하는 과정이 미분/적분 등에 유리하도록 하기 위해 제곱을 취함
- **표준편차란?**
    - 분산에 제곱근을 취한 값
    - 제곱근을 취한 값을 별도로 쓰는 이유는, 분산에서 제곱을 통해 측정 단위가 없어진 부분을 다시 복원하여 해석할 수 있기 때문
    - 예를 들어, 어떤 가구의 제작 사양 오차 (cm)에 대한 분산을 구하면 그 값은 cm 단위로 해석할 수 없으나, 표준편차는 동일한 cm 단위로 얼만큼 떨어져있는지 해석할 수 있음
    - 따라서, 평균에 표준편차를 더한 값으로 평균에서 표준편차만큼 떨어진 위치를 단위를 유지하면서 확인할 수 있음

▼ **정규분포 (Normal Distributiion)에 대한 이해**

- 종 모양으로 생긴 **대칭성**을 만족하는 확률분포

- 세상의 많은 현상들이 정규분포로 설명이 되며, 이해하기 쉬운 분포이기 때문에 통계학에서 아주 중요한 분포
- 종의 가운데가 평균값이고, 표준편차는 평균을 중심으로 대칭으로 1sigma = 34.1%, 2sigma = 13.6%, 3sigma = 2.1%, 3sigma 이상이 0.1%의 확률을 가짐
    - 정규분포 확률을 쉽게 외워두는 방법
        - 1sigma = 68.2%
        - 2sigma = 95.6%
        - 3sigma = 99.8%
- 통계학에서 가장 중요한 함수와 이론이 '정규분포'와 '중심극한정리'임을 기억하자.
    - 이 두 개로 많은 현상을 정규분포로 근사하여 쉽게 해석하고, 검증하며, 설명할 수 있기 때문

▼ **편포도(Skewness)에 대한 이해**

- 모든 '분포'가 정규분포처럼 대칭성을 가지지 않음 (**비대칭 분포**)
- 대칭이 아닌 경우, 확률이 한 쪽으로 치우치거나 규칙적이지 않은 확률을 가짐
    - 현실에서 주로 마주하는 분포
        - 꼬리가 있는 방향으로 좌/우로 명명
        - 좌측 편포 Left (Negative) Skew
            - 왼쪽에 이상치가 많은 경우
            - 사이트 방문에 대한 장애 비율 분포 (일반적으로 잘 접속하지만 장애가 나는 경우는 드문)
        - 우측 편포 Right (Positive) Skew
            - 오른쪽에 이상치가 많은 경우
            - 소득에 대한 분포 (일반적으로 저임금자보다 고임금자 수가 압도적으로 적음)

▼ **평균, 중간값, 최빈값에 대한 이해**

- 데이터의 중심 경향성을 요약한 지표
- **평균(mean)** = (중복을 포함한 모든 관측값의 합계 / 표본 수)
- **중앙값(median)** = 데이터의 관측값을 정렬한 뒤 그 중 가운데가 되는 값
    - 짝수인 경우 2개의 값이 발생 → 2개의 값에 대한 평균이 중앙값이 됨
    - 중앙값에 대한 흔히 하는 실수 → 차트/분포도만 보고 분포도의 중앙을 중앙값으로 정하는 것
    - 중복된 관측값도 계산에 포함되기 때문에 이를 놓칠 수 있음. 분포와 데이터를 동일하게 생각하기 때문
- **최빈값(mode)**
    - 데이터에서 가장 자주 나오는 값
    - 데이터나 분포함수에서 가장 많은 빈도로 나오는 값
    - 최빈값은 분포도의 가장 최대값에 위치함
- **평균, 중앙값, 최빈값의 관계**
    - 분포가 대칭인 경우 → 평균 = 중앙값 = 최빈값
    - 분포가 비대칭인 경우
        - 좌측편포 → 최빈값 < 중앙값 < 평균
        - 우측편포 → 평균 < 중앙값 < 최빈값

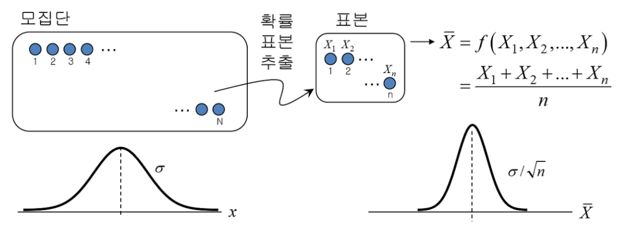
▼ **칸토어 대각선 논법을 통한 연속과 불연속에 대한 이해 (Optional)**

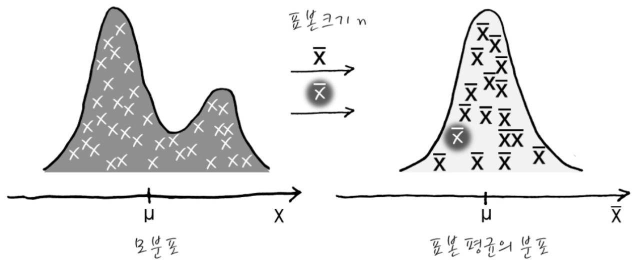
- 자연수 집합보다 실수 집합이 더 크다는 것을 증명한 논법으로, 0과 1 사이의 실수들이 자연수와 일대일 대응될 수 없다는 것으로 증명함
- 증명 방법
    - 0과 1사이의 모든 실수를
    - 자연수 1부터 무한하게 자연수에 대해 0과 1사이의 모든 실수를 임의로 1:1 대응시키는 리스트가 있다고 가정
        - 1번째 실수: 0.12345...
        - 2번째 실수: 0.67891...
        - 3번째 실수: 0.54321...
        - ...
    - 각 실수를 첫번째 실수일 때 첫번째 숫자를 가져오는 식으로 N번째 자리 숫자를 따와 새로운 숫자를 만드는 방법을 채택
        - 1, 2, 3번째 실수...를 통해 새로 만든 실수 = 0.173...
    - 이렇게 만든 새로운 실수는 대응 리스트의 모든 실수와 적어도 한 자리에서 다름을 찾아낼 수 있음
    - 따라서, 처음 가정했던 0과 1 사이에 모든 실수를 리스트에 나열할 수 있다고 가정했으나 그렇지 않은 반례가 나옴으로써 모순이 됨

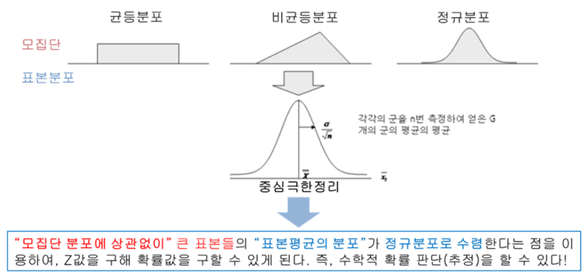
- 즉, 실수의 무한성은 자연수의 무한성과 다르다는 결론을 얻음

▼ **중심극한정리 (Central Limit Theorem) 의 이해**

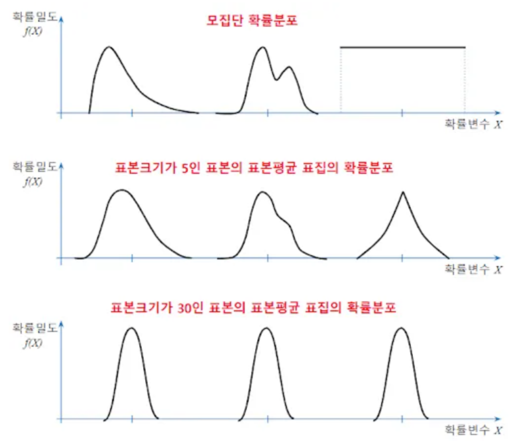
- **표본평균분포에 대한 CLT 만족 상황**







- **단일 표본에 대한 분포의 n= 30 이상일 때의 CLT 만족 상황**



> 직관적 정의: 모집단의 분포가 무엇이든지 상관없이 표본 크기가 충분히 크다면 표본분포가 정규분포에 근사하게 된다는 정리 (모집단과 같은 분포가 된다는 것이 아님!)

> 요약: 모집단의 분포가 무엇이든 상관없이 표본을 충분히 많이 추출할수록
> 1. 표본평균의 분포가 정규분포에 근사한다.
> 2. 표본평균의 표본분포의 평균은 모집단의 평균과 같아진다.
> 3. 표본평균분포의 표준편차는 모집단의 표본 크기 제곱근을 모집단 표본편차를 나눈 것과 같으며, 표본 크기가 커질수록 점점 작아진다. 즉, 분산이 작아지면서 변동성이 작아지고, 이에 따라 모집단의 평균에 더 가까워진다. (표본평균분포의 표준편차 = sigma / sqrt(n), n은 표본 크기)

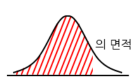
- **사전학습 - 모집단과 표본, 그리고 표본분포**
  - **모집단 (Population)** - 관찰할 수 있는 특정 사건에 대한 전체 집단
  - **표본 (Sample)** - 모집단에서 가져온 일부 집단
  - 모집단과 표본을 구분하는 이유
    - 모집단의 부분집합이 표본이므로, 모집단 전체를 설명하지 못함
    - 모집단을 완벽히 알기 어려운 현실적인 문제들이 대부분임
    - 파악할 수 있는 최대한의 한정된 표본을 가지고 전체 모집단을 추정하는 것이 통계의 관심 주제이므로 이를 구분해야 하고, 통계학 책에서도 별도의 notation으로 평균, 분산 등을 표기하는 이유도 마찬가지임
  - **표본분포 (Sampling Distribution)**
    - 모집단에서 무작위로 추출한 한 개의 표본에 대한 분포가 아님

- 모집단에서 무작위로 추출한 한 개 이상의 여러 개의 표본에 대한 통계량(주로 대푯값)의 분포
  - 표본평균에 대한 분포, 표본표준편차에 대한 분포 ...
- 표본평균분포는, 모집단에서 N개의 측정값에 대한 분포를 만들고, 그 분포의 평균을 계산하는 것을 반복하여 표본평균들을 가지고 분포로 만든 것

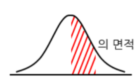
- **중심극한정리 (CLT)**
  - 표본을 추출할때마다 표본분포가 다양한 형태의 분포로 형성이 되면서 모집단의 실제 분포와는 다른 분포가 나올 수 있음
  - 하지만 계속해서 표본을 추출하면서 표본 추출하는 표본 크기가 크거나, 횟수가 높아질수록 표본분포는 정규분포의 형태로 점점 근사하게 되며, 이를 증명한 정리가 바로 중심극한정리
  - 직관적 예시
    - 어떤 책 하나의 모든 단어에 대한 분포가 정규분포가 아닐 수 있음
    - 하지만 (충분한)모든 페이지에 있는 단어 길이의 평균을 각각 구하여 표본평균의 표본분포를 만들면 그 표본분포는 정규분포가 됨
  - 직관적 아이디어
    - 모집단 분포에서 무작위로 표본을 계속하여 추출하여 분포를 그리고, 그것에 대한 표본평균을 구하면 쌍봉이거나 좌우측 편포가 되거나 다양한 모양의 분포의 평균을 계산하게 됨
    - 모집단 분포가 어떻게 되든 무작위로 표본을 추출하게 되면 모집단 평균에 비해 치우친 것과 중심에 가까운 표본평균에 대한 분포를 다양하게 얻지만, 모집단의 중심 경향성에 가까운 표본평균이 계산될 확률이 더 높음
    - 그래서 결과적으로 모든 표본평균에 대한 분포를 그리게 되면 정규분포에 근사하게 되고, 이 표본평균의 표본분포의 평균은 모집단의 평균과 표본 추출 횟수에 따라 점점 같아지게 됨
  - 내용 정리
    - 표본을 충분히 많이 추출할수록
    1. 표본분포가 정규분포에 근사한다.
    2. 표본평균의 표본분포의 평균은 모집단의 평균과 같아진다. (다른 대푯값/통계량에도 마찬가지)
    3. 표본분포의 표준편차가 작아진다. (표본표준편차 = sigma / sqrt(n))

▼ Z-Score 에 대한 이해

- **사전학습 - 정규화(Normalization)**
  - 일반적인 '정규화'의 의미는 데이터를 특정한 범위로 변환하는 방법
  - 통계학의 정규분포에서의 분포 정규화는 표준정규분포로 변환하는 'z-score 정규화' 를 의미함
- **Z-Score**
  - 데이터를 평균=0, 표준편차=1인 정규분포로 변환하는 정규화 방법
  - z = (x-mu)/sigma 로 z-score를 계산하며, x는 측정값, mu는 평균, sigma는 표준편차
- **표준정규분포**
  - 정규분포에 대한 z-score로 정규화하여 평균과 표준편차가 각각 0과 1인 정규분포로 범위를 변환시킨 것
  - 이 표준정규분포에 대한 특정 관측값 x에 대한 확률은 직접 계산하지 않고 미리 계산된 z-score 표를 활용하여 확인함
  - 표준정규분포표에서 관측값인 x = 0.~인 경우 z-score의 값이 0.0~인 경우의 셀에 있는 값이 확률값



Z	0.00	0.01	0.02	0.03
0.0	0.5000	0.5040	0.5080	0.5120
0.1	0.5398	0.5438	0.5478	0.5517
0.2	0.5793	0.5832	0.5871	0.5910
0.3	0.6179	0.6217	0.6255	0.6293
0.4	0.6554	0.6591	0.6628	0.6664
0.5	0.6915	0.6950	0.6985	0.7019
0.6	0.7257	0.7291	0.7324	0.7357
0.7	0.7580	0.7611	0.7642	0.7673
0.8	0.7881	0.7910	0.7939	0.7967
0.9	0.8159	0.8186	0.8212	0.8238



Z	0.00	0.01	0.02	0.03
0.0	0.0000	0.0040	0.0080	0.0120
0.1	0.0398	0.0438	0.0478	0.0517
0.2	0.0793	0.0832	0.0871	0.0910
0.3	0.1179	0.1217	0.1255	0.1293
0.4	0.1554	0.1591	0.1628	0.1664
0.5	0.1915	0.1950	0.1985	0.2019
0.6	0.2257	0.2291	0.2324	0.2357
0.7	0.2580	0.2611	0.2642	0.2673
0.8	0.2881	0.2910	0.2939	0.2967
0.9	0.3159	0.3186	0.3212	0.3238

▼ 중심극한정리에 대한 실전 비즈니스 과제1

- **문제 상황**
  - 해외 배송업체에서 일하는 분석가로, 빠르게 배달 가능 여부에 대해 경영진에게 알려줘야할 의무가 있음
  - VIP 고객이 요청한 바는, 현재 거주지에서 특정 지역까지 긴급하게 배송해줘야 하는 요청
- **사전 정보**
  - 고객이 이번에 적재하려고 하는 화물의 개수는 36개로 정해져 있음
  - 고객이 적재하려는 무게에 대해서는 유관부서에서 정확한 무게를 알려줄 수 없었음
  - 해당 VIP 고객의 이전 몇 년의 기록을 참고했을 때 거의 정확하다고 판단되는 값은
    - 이 화물은 평균 72파운드(32.66kg)
    - 표준편차는 3파운드(1.36kg)
  - 현재 화물을 실을 수 있는 비행기는 1개 뿐이고 총 2,630파운드(1,193kg)까지만 실을 수 있음

- **해결할 문제**
  - 모든 화물이 비행기에 안전하게 적재되어 운송될 확률은 얼마인지 알아내야함
  - 이를 통해, 화물 운송 가능 여부나 무게가 너무 많이 나가면 다른 방법을 찾아야한다는 점을 고객에게 빠르게 알려줄 수 있어야 함
- **문제 풀이**
  - VIP 고객이 이번에 요청할 수 있는 모든 가능한 화물을 모집단으로 함
  - 여기서 해당 고객이 이용한 모든 기록을 참고하여, 모집단의 평균이 72파운드, 표준편차는 3파운드임을 추정한 값을 알고 있음
  - 고객이 적재하려는 화물의 수도 36개로 정해져 있다는 것도 알고 있음, 즉 표본 크기는 고객이 보내주기로 한 36개가 표본이 됨
  - 하지만 현재 요청과 관련하여 고객이 얼만큼의 무게를 갖는 화물을 적재를 요청할지 모름 (즉, 표본을 고르는 건 고객이므로 모집단의 분포를 알 수 없음)
  - 일반적인 평균값을 이용하여 화물을 실을 수 있다고 말할 수도 있지만, 문제는 얼만큼의 확률로 화물을 싣고 이륙할 수 있는지임
  - 표본을 무작위로 추출할 수 있는 상황이 아니므로 (화물 무게에 대한 정보를 정확히 모르므로 표본을 많이 추출할 수 없는 상황) 이 36개의 표본에 대한 표본평균분포를 구했다고 가정하고, 이 것이 결과적으로 중심극한정리를 만족한다고 우선 가정한다.
  - 그러면, 표본분포의 평균 = 모집단 평균 = 72파운드
  - 표본표준편차 sigma는 모집단의 표준편차 / sqrt(36) = 3/6 = 0.5
  - 비행기에 실을 수 있는 총 화물 무게는 2630파운드
  - 고객이 실을 화물의 무게가 비행기 총 화물 무게를 넘어서면 이륙할 수 없으므로, 이륙할 수 있다는 가정하에 36개의 화물을 실으려면 고객의 요청사항에 대한 이륙확률의 위치는 다음과 같이 계산됨
    - x_crit = (2640 lb / 36 boxes) = 73.33 lb/box
  - 해당 위치에 대한 확률을 계산하기 위해 표준정규분포로 변환한다.
    - z-score = (x_crit-mu)/sigma = (73.33 - 72)/0.5 = 2.66
  - 표준정규분포표를 참고하면, P(x < x_crit) = 0.9961, 즉 99.61%
  - 그러면 이륙하지 못하는 확률은 100% - 99.61% = 0.39%
  - 따라서, 이륙 가능하다고 답변드릴 수 있고, 그 확률이 99.61%이라고 말씀드릴 수 있음 (또는 이륙 못할 리스크를 0.39% 감수해야 함)

▼ **중심극한정리에 대한 실전 비즈니스 과제2**

- **문제 상황**
  - 월스트리트 투자 펀드에서 퀀트로 일하고 있음
  - 트레이더 팀에 대한 supervisor로써 많은 트레이더를 관리하고 있음
- **사전 정보**
  - 수익은 '라플라스 분포'에 근접하게 얻고 있음
  - 수익의 평균은 95.7달러이며, 1,247달러의 표준편차를 갖고 있음
  - 매주 약 100건의 거래를 함
- **해결할 문제**
  - 문제A : 내 팀이 한 주에 손해를 볼 확률은?
  - 문제B : 한 주에 2만 달러 이상의 수익을 얻을 확률은?
  - Hint: 어떤 분포에서 수익이 발생하는지가 중요할까요? → NO, 아무 상관 없음. 중심극한정리는 모집단의 분포가 무엇인지 상관하지 않음
- **문제 풀이**
  - 수익에 대한 모집단 분포는 라플라스 분포임을 알고 있고, 이에 대한 평균이 95.7달러, 표준편차는 1,247달러임을 알고 있음
  - 매주 약 100건의 거래를 한다는 점을 알고 있으나, 이는 전체 모집단에 대한 거래량은 아니므로 표본 크기로 설정할 수 있음
  - 매주마다의 표본을 추출하여 충분한 표본평균분포를 계산한다고 가정하고, 이에 대해 중심극한정리의 3가지 특징을 적용하면 다음과 같음
    1. 수익에 대한 표본평균분포 = 정규분포
    2. 1번에 의거, 표본평균 = 모집단 평균 = 95.7달러
    3. 표본표준편차 sigma = 모집단표준편차 / sqrt(100건) = 1,247 / 10 = 124.7 달러
  - 이제, 표본평균분포 = 정규분포이므로, z-score를 계산하여 scaling한 뒤, 문제 A와 B에 대한 확률을 계산한다.
  - 문제 A에 대한 확률은 표준정규분포표에서 x ≤ z인 확률 중 우측에 대한 표를 이용하였음. 즉, z-score = (x-mu)/sigma 일 때 x가 0보다 낮은 값이므로 (0-95.7)/124.7 = -0.77 이고, 우측 기준 x ≤ z일 확률은 P(x ≤ z) = 77.94% 손해를 보는 쪽이므로 평균보다 낮은 값인 좌측 분포의 확률을 구해야하므로, 정규분포가 대칭인 점을 이용하여 최종적으로 100%-77.94% = 22.06%
    - 한 주에 손해를 봐야 하므로 수익이 0달러보다 같거나 작아야하고, 직관적으로는 0달러가 되면 수익이 없는 것이므로 손해라고 봐야함. 그러므로 한 주 100건에 의해 건당 수익도 0달러
  - 문제 B에 대한 확률은 표준정규분포표 기준으로는 한 주에 2만 달러 미만을 얻을 확률을 전체 확률에서 뺀 것과 같으므로, z-score = (200달러 - mu)/sigma = 0.84 ⇒ P(x ≤ z) = 79.95% ⇒ 100% - 79.95% = 20.05%
    - 한 주에 20,000달러의 수익을 얻으려면 100건의 거래건수에 대해 평균 200달러의 수익을 얻어야 함. 즉, 20,000달러 / 100건 = 200달러/건

▼ **통계적 유의성 (Statistical Significance) 에 대한 이해**

> 통계적 유의성은 귀무가설이 참이라 가정하고 귀무가설을 기각할 유의확률을 신뢰수준에 따라 결정하는 것

- **사전 학습 - 귀무가설과 대립가설**
  - 귀무가설: 모집단 간 차이가 없다 / 공정하다
  - 대립가설 : 모집단 간 차이가 있다 / 불공정하다
- **유의확률 (p-value)**
  - 귀무가설이 참인 집단에서 귀무가설이 참이 되지 않을 확률
- **신뢰수준 (confidence level, alpha)**
  - 귀무가설이 참인 집단에서 대립가설이 참이거나 불공정한 결과가 우연일 가능성이 낮다고 보는 유의확률의 기준을 세우는 값
  - 일반적으로 0.05 (5%)로 설정하나, 생명과 관련된 의료 실험 등에서는 1%로 설정하기도 함

▼ **가설검정 (Hypothesis Testing) 에 대한 이해**

> <u>간단 요약:</u>
>
> 1. 가설검정은 모집단 간 차이가 없다는 귀무가설을 참으로 가정하고,
> 2. 주어진 표본을 통해 계산된 표본분포의 통계량이 귀무가설을 기각할 만큼 통계적 유의성이 확보된다면
> 3. 이를 기각하여 모집단 간 차이가 있다고 말할 수 있는(대립가설을 채택하는) 통계적 방법론
> 4. 주로 평균에 대한, 때로 비율에 대해서도 모집단 간 가설검정을 실시할 수 있음

- 일반적 가설검정은 중심극한정리를 이용한 확률 계산에 기반을 두고 있음
- 그러므로 중심극한정리 속 숨겨진 조건들을 만족하는 것이 중요함
- <u>z-score를 활용하는 Z-test 가설검정 조건이 가설검정하기 가장 쉽고 이상적인 조건으로, 이를 늘 가정할 수 있는지 확인해야함</u>
  1. 표본은 무작위로 선택되어야 한다.
     - 특정 경향이나 행동을 한 사람이나 사건이 매우 많이 들어가있다면 데이터는 편향된 것
     - 무작위의 의미가 없어 CLT를 적용할 수 없고, 표본크기를 키우는 것도 의미가 없음
  2. 관측값은 서로 독립이어야 한다.
     - 관측되는 대상/사건이 서로 영향을 미치는 관계라면 편향될 가능성이 높음
  3. 모집단의 표준편차가 알려졌거나, 표본이 적어도 30개 이상이어야 한다.
     - 중심극한정리가 작동할 가능성이 높은 최소 표본 개수가 n = 30 이상
     - 만약 모집단의 표준편차를 모르거나 이 조건을 만족하지 못하면 스튜던트 T-검정을 사용해야함

▼ **가설검정 연습문제 풀어보기 (Z-검정)**

- **문제 상황**
  - 숟가락 제조 공장에서 1000만 달러를 투자해서 23%였던 제조상 불량률로 인한 높은 반품률을 방지하고자 장비와 제조 과정을 업그레이드함
- **사전 정보와 문제 요청**
  - 무작위 표본으로 150개의 숟가락을 받았고 그 중 23개의 숟가락이 불량이었음
  - 그래서 95% 신뢰수준으로 새로운 장비로 상황이 개선되고, 불량 숟가락의 수가 18% 이하로 감소한 것을 증명해달라고 요청함
- **문제 풀이**
  1. 가설 설정

     H0: 업그레이드 이전과 이후의 불량률에는 차이가 없거나 더 크다. (p ≥ 18%)

     H1: 업그레이드 이전에 비해 이후의 불량률이 감소했다. (p< 18%)
  2. 사전 정보 체크
     - 모비율 = 0.18, 1-모비율 = 0.82
     - 표본 크기 n = 150, 신뢰수준 = 0.05, 불량 수 = 23
     - 표본비율 = 23/150 = 0.1533
     - 표본크기가 n=150이므로 중심극한정리에 의해 정규분포를 근사적으로 따름을 가정
  3. 비율 검정 시 체크할 것
     - 두 반대되는 비율이 일정량의 표본 (10 이상)을 확보해야 한다는 조건
     - 이를 만족하지 않으면 표본 수를 더 늘려야함
     1. n * p = n * 모비율 > 10

        - 150 * 0.18 = 27
    2. n * q = n * (1-모비율) > 10
        - 150 * 0.82 = 123
    - → 모두 만족함
4. 통계량 계산
    - 표본평균 = 모평균 = 모비율
        - 0.18
    - 표준편차 = sqrt(pq) (모비율과 1-모비율 곱의 제곱근)
        - sqrt(0.1533*0.8467) = 0.384
    - 표본비율 = 0.153
    - 표본표준편차 = 모표준편차/sqrt(표본크기) = 0.031
5. z-score 계산
    - z = (표본비율 - 모비율) / 표본표준편차 = (0.153 - 0.18) / 0.031 = -0.87
    - 표준정규분포표에 따르면, P(x ≤ z) = 80.78%
    - 표본비율은 모비율보다 작으므로 정규분포의 대칭성을 이용해 반대측의 확률로 계산해야함
    - 즉, 이 z-score에 해당하는 유의확률(p-value)는 100%-80.78% = 19.22%
6. 통계적 유의성 체크 (2가지 방법)
    - p-value를 계산하여 비교
        - 신뢰수준이 95%로 설정되었으므로, 5%보다 작아야하며, 유의확률이 19.22%로 5%보다 큼
    - 기각역을 이용한 방식
        - 유의수준에 대한 z확률을 구하고 검정통계량과 유의수준에 따른 z확률을 비교하면 됨
7. 결론
    - p-value가 19.22%로 95% 신뢰수준에서, 다른 말로는 5%의 유의수준에서 귀무가설을 기각할 수 없음
    - 즉, 업그레이드 이전의 불량률 23%에서 업그레이드 이후 불량률이 18% 이하로 감소하였다고 볼 수 있는 증거가 통계적으로 불충분하다고 말할 수 있다.
    - (= 기존과 유의미한 차이가 있다고 말할 수 없다.)
    - 단, 'H0을 기각할 수 없다고 해서 대립가설이 참인 것인지와 귀무가설이 틀렸다고 말할 수 없다. 그저 대립가설을 채택할 증거가 더 필요하다는 것'

▼ 귀무가설을 기각할 수 없다는 말이란?

> 귀무가설에 모순이 발생하지 않은 경우, 귀무가설을 채택하거나 귀무가설이 옳다고 말하는 것이 아니라, 귀무가설을 기각할 증거가 충분하지 않다는 것, 즉 증거가 더 모이면 귀무가설을 기각할 수도 있다는 것.

- 귀무가설을 기각할 수 없지만 동시에 대립가설도 기각할 수 없음
- 귀무가설이 기본 시나리오이므로 귀무가설만 기각할 수 있음
- 귀무가설이 모순에 도달하지 않으면 귀무가설이 옳지 않다는 뜻, 또는 귀무가설을 기각할 증거가 충분치 않다는 뜻
- 만약 귀무가설에 모순이 발생했으면 대립가설을 채택함으로써 귀무가설을 기각할 수 있음
- 하지만 귀무가설에 모순이 발생하지 않은 경우, 귀무가설을 채택하거나 귀무가설이 옳다고 말하는 것이 아니라, 귀무가설을 기각할 증거가 충분하지 않다는 것, 즉 증거가 더 모이면 귀무가설을 기각할 수도 있다는 것.

▼ t-분포 (Student's t-Distribution) 에 대한 이해

> t-분포 특징 요약:
>
> - 표본 크기가 적을 때 사용하기 위해 고안된 분포
> - 정규분포와 아주 비슷하나, '자유도'에 따라 모양이 달라짐
> - 자유도가 증가한다는 것은 표본 크기가 증가한다는 것
>   - 자유도 = 표본크기 - 1
>   - 자유도가 높아질수록 정규분포에 가까워짐
>   - 자유도가 낮아질수록 꼭대기가 낮아지고 꼬리 부분이 두꺼워지면서 정규분포와 다르게 이상치도 활용될 가능성이 있음

- **t-분포를 사용하는 경우**
    1. 모집단 표준편차를 알 수 없으며

2. 표본 크기가 작은 경우 (n < 30)
    - 단, 표본 크기가 30 이상이면 정규분포와 거의 동일해짐

- t-분포의 검정통계량
    - t = (x_bar - mu) / (s/sqrt(v))
    - (표본평균과 모평균 차이) / (표본표준편차 / 자유도의 제곱근)
    - 따라서 z-분포와 다르게 모표준편차를 알지 못해도 상관없음

▼ **T-검정에 대한 이해**

- 기본적으로 Z-검정과 동일한 원리로 가설검정
- **Z-검정과의 차이**
    - 단, t-분포에서의 사용 조건과 자유도, 검정통계량 계산에 대해 숙지해야함
        - 표본크기가 30 이하로 작을 때
        - 모표준편차를 모를 때
        - 자유도는 표본크기 - 1
        - 검정통계량에서는 표본표준편차를 사용하고, 표본크기가 아닌 자유도의 제곱근을 나눔
    - T-검정에서는 Z-검정처럼 z-score로 변환한 검정통계량을 가지고 확률을 바로 알아낼 수 없음
        - 자유도와 신뢰수준을 정해야하기 때문

▼ **단측검정과 양측검정에 대한 이해**

> 결과의 방향이 불확실하거나 관련 사전 지식이 없을 때 양측검정을 적용해서 더 엄격하게 검정할 수 있음

- **단측검정 (One-Tailed Test)**
    - H0 = H1이나 H0 ≥ H1, H0 ≤ H1 을 가정하고, H0 > H1 또는 H0 < H1 를 채택하려는 것
    - 신뢰수준 95% 기준 한쪽 꼬리에 5%의 기각역
- **양측검정 (Two-Tailed Test)**
    - H0 = H1을 가정하고 H0 ≠ H1 을 채택하려는 것
    - 신뢰수준 95% 기준 양쪽 꼬리의 2.5%의 기각역
    - 기각역이 더 좁아지므로 귀무가설을 기각하기 더 어려워짐
    - 이 때문에 양측검정이 더 엄격한 검정
- 양측검정이 사용되는 예시
    - 의약품 출시를 위한 가설검정
    - 약이 효과가 있었을 수 있지만 오히려 부작용이 발생할 수도 있기 때문
    - 단측검정에서는 사전 지식이 있다고 가정하기 때문에 나머지 한쪽은 고려대상이 되지 않음
    - 따라서, 사전 지식이 확실하지 않다면 양측검정을 이용하여 더 엄격하게 가설을 검정하고자 함
    - 실제 계산은, 유의수준을 반으로 나눠서 똑같이 검정하면 됨

▼ **유의확률 오용과 모호성에 대한 이해**

- 약을 구매하면 약의 부작용/효과에 대해 검정을 통해 임상 증명된 약임을 홍보함
- 유의확률이 5%미만, 1%미만인 결과를 증명함
- 하지만 규모에 대해서는 언급하지 않음
- 규모가 중요한 이유는, 규모가 달라짐에 따라서 가설검정의 결과가 달라질 수 있기 땨문임
- 가설이 틀리게 검정된 것은 아니지만, 규모와 같은 핵심적인 정보들을 자세하게 들여다봐야 한다는 이야기