파이썬 라이브러리를 활용한 데이터 분석

툴 스터디 3주차

지난주!

- EDA 과정과 방법을 이해해 보았습니다.
- 바 그래프를 그리며 데이터를 시각화해 보았고,
- 평균, 중앙값, 최빈값, 범위, 분산 등의 통계 지표를 박스 플롯 차트를 통해 이해하였습니다.
- 상관관계의 개념을 이해하고, heatmap 차트를 그려 타이타닉 데이터 변수 간의 상관관계를 확인하였습니다.

오늘은?

- 데이터 분석은 우리가 입증하고자 하는 가설을 세워 데이터 분석을 통해 가설이 합당한지, 그렇지 않은지 확인하기 위해 진행합니다.

- 오늘은 기초적인 확률·통계 개념을 배운 다음,

  개념을 바탕으로 데이터 분석 과정의 핵심인 '추론'과 '가설 검정' 과정을 이해해보겠습니다!

확률이란?

- **확률(Probability):** 어떠한 '사건의 공간' 에서 '특정 사건'이 발생할 가능성을 '수치'로 나타낸 것

  Ex) 예를 들어, '하나의 동전을 던지는' 사건에서

  '동전의 앞면이 나올' 확률은 $\frac{1}{2}$

- 동전의 앞면이 나올 사건을 'E'라고 할 때,

  동전의 앞면이 나올 확률은 P(E)로 표기합니다.

종속성과 독립성, 조건부 확률

- 사건 A와 사건 B가 있을 때,
  사건 A의 발생 여부가, 사건 B의 발생 여부에 대한 '정보를 제공'한다면,
  사건 A와 사건 B는 **종속 사건(Dependent Event)**

- 사건 A와 사건 B의 발생 여부가 서로 상관이 없을 때,
  사건 A와 사건 B는 **독립 사건(Independent Event)**

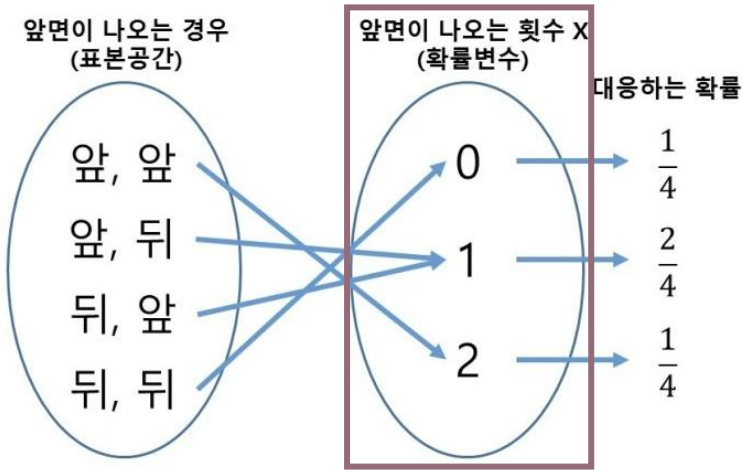
- 사건 A와 B가 동시에 발생할 확률 **P(A, B) = P(A)P(B)**일 경우, 사건 A와 B는 **독립 사건!**

- **조건부 확률(Conditional Probability):**
  P(B|A) = $\frac{P(A \cap B)}{P(A)}$, 사건 A가 일어났을 때, 사건 B가 발생할 확률

확률 변수와 확률 분포

- **확률 변수 (Random Variable)**

  특정 값이 나타날 가능성이 '확률적으로 주어지는' 변수

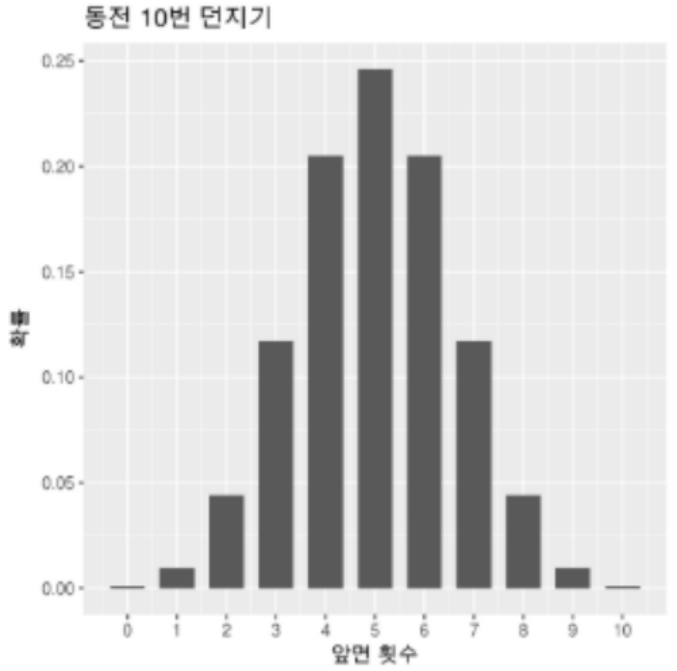


- **확률 분포 (Probability Distribution)**

  확률 변수가 특정한 값을 가질 확률을 의미합니다

  오른쪽 차트와 같이 각각의 확률 변수의 확률 분포를

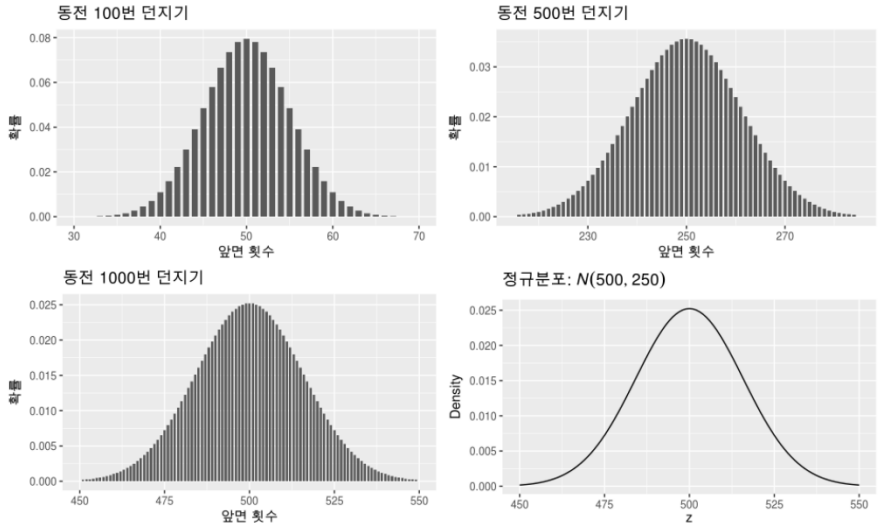
  나열한 것을 우리는 **확률분포표** 라고 부릅니다.



정규분포와 표준 정규 분포

- **중심극한정리:** 서로 독립 관계를 가진 **확률 변수 n개의 평균의 분포**는

  n이 적당히 크다면 **정규분포에 근사함** (그래프 형태가 정규분포에 가까워 짐)



- **정규분포(Normal distribution):** 평균을 중심으로 그래프의 **좌측과 우측이 대칭인** 확률 분포

- **표준 정규 분포(Standard normal distribution):**

  정규 분포를 표준화하여 만들어진 **평균 = 0, 표준편차 = 1**인 정규 분포를 의미합니다.

정규분포와 표준 정규 분포

- **정규화 (Normalization):** 여러 대상들을 일정한 규칙이나, 기준에 따르는 '정규적인' 상태들로 바꾸는 과정

- **표준화 (Standardization)를 하는 두가지 이유**

  1) 서로 다른 자료를 비교, 분석할 수 있게 됨

  Ex) 국적과 통화가 서로 다른 두 회사원의 연봉 평균을 표준화하여 둘의 연봉 평균을 비교할 수 있게 됨

  2) 표준 정규 분포표를 이용해 근사값을 찾기만 하면 **복잡한 확률의 계산을 간편하게** 할 수 있음

추론

- 모집단에서 추출한 **표본**을 가지고 **모집단의 특성을 추정**할 수 있습니다.
- 그리고 그 결과가 신뢰성이 있는지 검정하는 과정을 통틀어 **추론**이라고 합니다.

- 추론 과정에서 우리는 다음과 같은 질문의 답을 얻어낼 수 있습니다.
- **1. 표본집단이 모집단을 대표할 수 있는가?**
- **2. 표본의 확률분포는 어떠한가?**
- **3. 추정된 결과는 신뢰성이 있는가?**

가설 검정

- **우리는 표본을 통해 모집단에 대한 주장(가설)의 옳고 그름을 판정할 수 있습니다.**
- 그리고 가설 검정을 통해 **대립가설(H1)을 채택하고, 귀무가설(H0)을 기각**합니다.

- **귀무가설은 처음부터 '기각할 것'을 가정하고 설정하는 가설**입니다.

  귀무가설은 **'ㅇㅇ과 ㅇㅇ은 차이가 없다', 'ㅇㅇ과 ㅇㅇ은 같다',**

  'ㅇㅇ은 ㅇㅇ에 영향을 미치지 않는다' 등으로 설정합니다.

- **대립가설은 우리가 '옳음'을 확인하고자 하는 가설이며,** 귀무가설과 대비되도록 **'ㅇㅇ는 ㅇㅇ보다 많다', 'ㅇㅇ은 ㅇㅇ보다 적다', 'ㅇㅇ은 ㅇㅇ에 영향을 준다'** 등으로 설정합니다.

- 통계적 가설 검정의 궁극적인 목표는 **대립가설이 옳은 지 확인**하는 것입니다.

  그렇기 때문에 처음에 귀무가설을 설정한 다음, 추론과 검정 과정을 통해 귀무가설을 기각하고,

  대립가설이 옳음을 확인하여, 대립가설을 채택하게 됩니다.

가설 검정

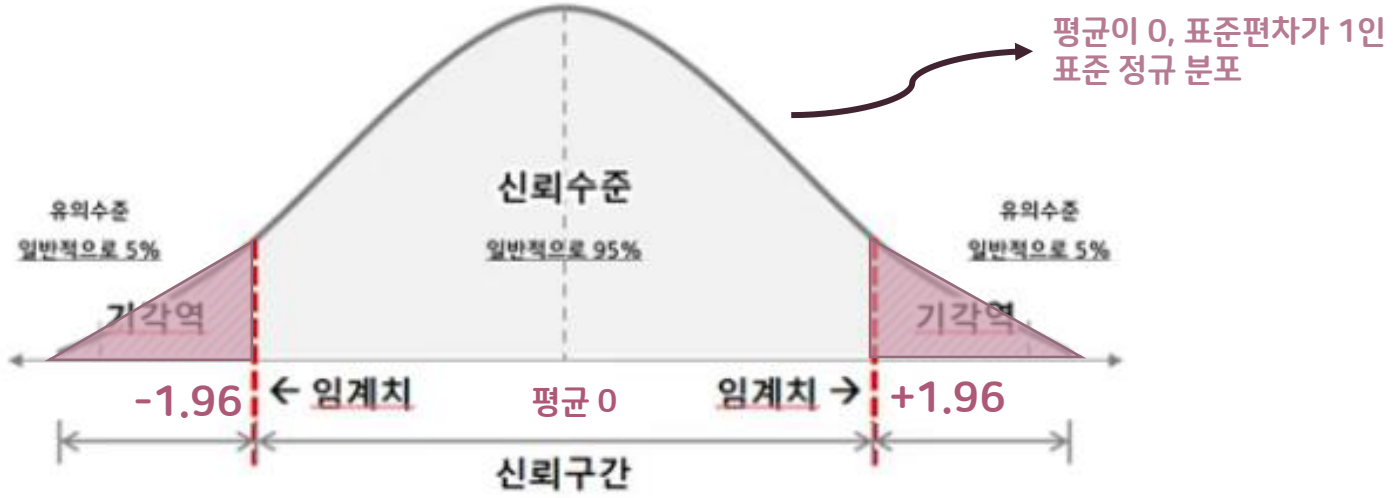
- **검정통계량 Z:** 표본 데이터를 기반으로 계산되어, **가설 검정에 사용**되는 랜덤 변수입니다.

  설정한 임계치를 기준으로, 검정통계량 값이 기각역 안에 속할 때,

  귀무가설을 기각합니다.

- **신뢰수준:** 일반적으로 **95%**, 혹은 **99%**로 설정합니다.
- **유의수준:** 신뢰수준이 95%일 때, **5%** | 99%일 때, **1%**입니다.
- **임계치:** 신뢰수준이 95%일 때, **-1.96, +1.96** | 99%일 때, **-2.58, +2.58**
- **'표본'이 신뢰수준 95%를 벗어날 경우 (= 검정 통계량 Z 값이 유의수준 5%에 속하여, 기각역 안에 포함될 경우)**

  귀무가설이 틀렸음이 인정되고, 귀무가설을 기각하고 대립가설을 채택할 수 있다.



가설 검정

- 검정통계량 Z 값 공식: $\dfrac{\text{표본평균} - \text{모평균}}{\dfrac{\text{표준편차}}{\sqrt{n}}}$

$$Z = \frac{\overline{X} - \mu}{SE}$$

$\overline{X}$ 표본평균,μ는 모 평균, SE는 표준오차

$$SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

σ는 모집단 표준편차(standard deviation), n은 모집단의 크기

가설 검정 과정 실습

- 가설 검정 과정

- 1. 가설 설정 (귀무 가설과 대립 가설)
- 2. 유의 수준 설정
- 3. 검정 통계량 산출
- 4. 가설 기각/채택 판단

가설 검정 과정 실습

- 전체 인구 IQ의 평균은 100이고, 표준편차가 15입니다.
  연구팀이 개발한 신약이 인간의 IQ에 영향을 주는지 알아봅시다.

- 1. 귀무가설(H0)과 대립가설(H1) 설정:
  귀무가설: 신약을 복용한 사람들의 IQ 평균은 100이다. (신약은 IQ 변화에 영향을 주지 않는다)
  대립가설: 신약을 복용한 사람들의 IQ 평균은 100이 아니다.
  (신약 복용이 IQ 변화에 영향을 준다)

- 2. 유의수준 설정:
  유의수준을 5%로 설정 (= 검정통계량 Z값이 신뢰수준 95%를 벗어난다면 귀무가설 기각)
  5%이므로 유의수준은 0.05!

가설 검정 과정 실습

- 가설 검정을 위해 모집단에서 실험 참여자 30명을 표본으로 추출하여 이들의 IQ를 샘플링하였습니다. 샘플링한 표본의 평균은 140이었습니다.

- 3. 검정통계량 산출:

  표본평균: 140, 모평균: 100, 표준편차: 15, $\sqrt{n}$: $\sqrt{30}$

  $$Z = \frac{140-100}{\frac{15}{\sqrt{30}}} = \frac{40}{2.74} = 14.6$$

가설 검정 과정 실습

- 4. 가설 기각/채택 판단

  유의수준: 0.05 / 임계치: -1.96, +1.96 / 검정통계량 Z: 14.6

  귀무가설 '이 약을 복용한 사람의 아이큐 평균은 100이다.'는 기각될까요?

  => 검정통계량 Z 값이 임계치 1.96보다 크므로, 표본은 신뢰수준 95%를 벗어나 기각역에 포함됩니다!

  따라서 귀무가설을 기각하고, 대립가설을 채택할 수 있습니다.

  신약을 복용한 사람의 IQ 평균은 100이 아니다 (신약은 IQ의 변화에 영향을 준다)라는 대립가설은 유의수준 0.05에서 통계적으로 유의미하다고 결론을 내릴 수 있습니다.

다음주!

- 다음주 툴 스터디는 팀장 개인 일정으로 인해 녹화 영상으로 제공될 예정입니다. 😭
- 다음주에는 쥬피터 노트북을 이용해 오늘 배운 추론과 가설 검정 과정을 실습해볼 예정입니다!
- 그리고 간단한 웹 크롤링 개요를 설명 드리도록 하겠습니다 😊