통계 정리

B2_김지현

기술통계

(표본) 데이터의 속성을 특정 통계량을 사용해 정리, 요약, 설명하는 방법

중심 척도

중심 경향성(중심적인 경향을 나타냄)을 보임. 산포 정도까지는 설명하지 못함

1. 산술평균 Mean

- 관측치의 총합을 관측치 개수로 나눔
- 가장 보편적이고 대표적인 대표값임
- 이상치에 민감
- 수치 척도에 의미가 있지만, 순서 척도인 경우에도 사용

2. 중위값 Median

- 전체 주어진 관측치를 크기 순으로 나열했을 때, 중앙에 위치하는 관측치
- 이상치/특이치에 영향을 덜 받음(실데이터에 의존하지X 때문)
- 데이터의 개수가 짝수면, 가운데 2개의 데이터를 평균 낸 값이 중위값

3. 최빈치(값) Mode

- 이산형 데이터에서 사용(value_counts)
- 전체 주어진 관측치들 중 가장 빈도가 높은 값으로 정의함

산포 척도 Degree of Dispersion

데이터의 퍼져 있는 정도를 설명하는 기술 통계

1. 분산 Variance

- 모분산: 편차 제곱의 합을 표본의 개수로 나눠줌
- 표본분산: 편차 제곱의 합을 (표본개수-1)로 나눠줌

2. 표준 편차 Standard Deviation

- 분산에 제곱근을 하여 구한 값
- 특이치에 민감하여 영향을 받음!!!!!!!!!

3. 사분위수 범위 IQR

- 사분위범위 = 3분위수(75%) - 1분위수(25%)
- 범위에 비하여 이상치의 영향을 덜 받음(사분위수 바깥에 위치한 특이값을 고려하지 않기 때문)
- 이상치 판단의 기준이 됨: Q3 + 1.5IQR, Q1 – 1.5IQR

4. 범위 Range

- 최대치, 최소치의 차이 = 범위 = 최대값 – 최소값
- 범위가 넓다 = 데이터가 퍼져 있다.

분포 모양

데이터가 퍼져 있는 형태를 나타냄

1. 도수 분포

- 도수: 구간 별로 x 데이터 개수를 셈
- 상대도수: 비율
- 누적도수: cumulative

2. 비대칭도(왜도)

- 분포가 어느 한쪽으로 치우친 정도를 나타냄
- left skewness(= negatively skewed) 왼쪽으로 꼬리가 길게 빠짐(Mode가 오른쪽에 위치)
- right skewness(= positive skewed) 오른쪽으로 꼬리가 길게 빠짐(Mode 왼쪽에 위치)

3. 첨도 Kurtosis

- 분포 모양이 얼마나 뾰족한가를 나타내는 정도
- 0과 같으면 정규분포와 같은 형태, 0보다 크면 뾰족하고, 0보다 작으면 완만한 형태

자주 보는 통계 기호

- 모평균, 모분산, 모표준편차, 모집단의 상관계수(ρ), 모수(θ)는 모두 상수임(변하지 않는 값)
- 표본평균, 표본분산, 표본표준편차, 표본상관계수(r)은 모두 변수임(변할 수 있는 값)

- 유의수준 알파: 신뢰수준 + 유의수준 = 1
- 회귀계수, 오차(실측값 – 예측값)(y = y hat)

데이터 유형

연속형 데이터

- 나누어질 수 있고, 연속적으로 측정될 수 있는 것
  예: 무게, 온도, 강도 등 계량형 데이터 (측정 단위들이 있음)
- **등간 척도**
  - 같은 간격을 가짐. 그러나 수치적으로 비율관계가 성립되지 않음
  - 즉, 크기의 차이가 절대적이지 않고 상대적임
  - 등간 척도로 측정된 변수들 간에는 가감(+-) 연산이 가능
  - 온도, 물가지수 등이 그 예임
- **비율 척도**
  - 등간 척도에 비율의 개념이 추가됨
  - 비율로서 표시하기 때문에 절대적 기준값이 존재함
  - 가감뿐 아니라 모든 산술적 사칙연산이 가능함
  - 중량, 강도 등이 그 예임

이산형 데이터(범주형 데이터)

- 나누어질 수 없고, 발생 빈도(count)를 세어서 countable 산출
  예: 적합/부적합, 1등급/2등급/3등급 등 (셀 수 있음)
- **명목 척도**
  - 관찰대상의 속성에 따라 관찰대상을 상호 배타적이고 포괄적 범주로 구분하는 데이터
  - 변수 간 사칙연산이 의미 없음
  - 성별, 품질(양품, 불량), 운동선수 등번호, 종교 등 범주형으로 구분되는 것이 그예
- **순위 척도**
  - 관찰대상이 가지는 속성 크기에 따라 순위, 서열을 부여함
  - 만족도, 성적등급, 크기 등이 그 예

cf. 연속형 데이터와 이산형 데이터는 분석 방법이 서로 다름

데이터 유형에 따른 통계분석 방법 및 대표값

데이터 유형		분류	순위	동일 간격	절대영점	대표값	통계분석
이산형	**명목척도**	O	X	X	X	최빈값, 퍼센트	빈도분석, 비모수
	순위척도	O	O	X	X	중위값, 퍼센트	비모수
연속형	**등간척도**	O	O	O	X	산술평균	모수통계
	비율척도	O	O	O	O	산술평균, 기하평균	

- 분류는 모든 데이터 유형에 대해 가능하다.
- 순위를 매기는 것은 명목척도만 불가능하다(봄 여름 가을 겨울에 서열은 없음)
- 동일간격은 연속형(수치형) 데이터만 가능하다.
- 절대영점은 연속형 중 비율척도만 가지고 있다.
- 이산형 공통: 퍼센트
  명목: 최빈값 vs. 순위: 중위값
  비모수 통계 활용, 명목척도는 빈도분석 가능(봄이 얼마나 많이 선호되는가? 등)
- 연속형 공통: 산술평균
  등간: - vs. 비율: 기하평균
  모수 통계 활용

확률분포

확률변수가 특정 값을 가질 확률, 즉 상대적 가능성을 나타냄

모든 확률변수(X) 값과, X가 발생할 확률 값을 도수분포표 혹은 그래프로 나타낸 것

확률

경험 혹은 실험의 결과로 특정사건이나 결과가 발생할 가능성을 의미

- Random Process: 어떤 결과들이 발생할지는 알고 있음. 그러나 어떤 결과가 나올지는 모름
- Mutually Exclusive Outcome: 동시에 두 가지 사건이 발생할 수 없음
- Law of Large Number: 시행 횟수가 많아질수록 특정 결과가 발생할 비율은 자체의 가능성에 점점 수렴함
- Independence: 한 시행의 결과는 다른 시행의 결과에 영향을 주지 않음

확률변수

표본공간 상 가능한 모든 결과에 특정 수치를 부여 (ex. 동전을 던져 앞면이 나오는 횟수)
cf. 표본공간 Sample Space: 실험 결과 발생할 수 있는 모든 가능한 결과의 집합

확률실험의 결과를 실수에 대응시키는 함수(또는 방법)

- **이산확률변수**
  - 확률변수가 취하는 값이 유한개
  - 이산확률함수(PMF. 확률질량함수): 확률변수의 확률 값을 모두 더하면 1
- **연속확률변수**
  - 구간 내 임의의 모든 점을 취할 수 있을 때
  - 연속확률함수(PDF, 확률밀도함수): 확률함수의 면적은 1

cf. 확률함수: 확률변수 X에 대하여 정의된 실수를 0~1 사이의 실수(확률)에 대응시키는 함수.
확률변수의 분포를 나타낸다.

확률 계산 (41 페이지 참고, 중요개념)

1. 확률밀도 Probability Density (연속확률함수)

- X에서 확률밀도함수의 값 f(x)

2. 누적확률 Cumulative Probability

- P(X ≤ x)의 값
- 확률변수의 범위가 음의 무한대에서 특정 X 지점까지로 정해지고, 해당 지점까지 확률밀도함수의 누적 면적을 의미
- x가 주어졌을 때, 해당 x까지의 누적 y값

3. 역 누적확률 Inverse Cumulative Probability

- 누적 y값이 주어졌을 때(연속확률함수, 확률밀도함수의 누적면적), 해당 확률 값에 대응하는 x 값 (확률을 알 때 해당 x 값)

확률분포의 종류(데이터 유형에 따른 구분)

연속확률분포

- **정규분포(= 가우스 분포)**
  - 평균을 중심으로 좌우 대칭됨 (종 모양)
  - 측정 기준치와의 차이를 나타내는 측정오차가 어떤 특성을 갖는 분포 형태를 이룸
  - 평균은 정규분포의 위치를 결정하고, 분산은 정규분포의 모양을 결정함(교재 39페이지 참고). 데이터의 개수에 모양의 영향을 받지 않음.
    → 따라서 구간 별로 확률 값이 미리 계산되어 정해져 있음(1σ = 0.341, 2σ = 0.136)
  - 정규분포의 곡선 아래의 면적은 1임. 대칭이기 때문에 평균을 중심으로 0.5씩 면적이 나뉨.
  - 수집된 자료의 분포를 근사하는 데 자주 사용됨(ex. 이항분포).
    (중심극한정리 때문. 독립적 확률변수들의 평균은 정규분포에 가까워지는 성질이 있음)
  - 모수(모표준편차)를 알 때 모평균을 검정하고자 하면 정규분포를 이용함.
  - 장점: 범용적으로 쓰이고 있음 (통계 분석의 기본전제가 되는 경우가 많음)
    단점: 모집단의 특성(평균, 분산)을 알아야만 사용할 수 있음(모를 시 t분포)

  cf. 정규분포를 기반으로 대부분의 확률분포들이 파생됨

- **표준정규분포**
  - 확률변수 X를 Z로 정규화하여 평균 0, 표준편차 1로 만듦

- 다시 말해, 평균이 0이고 표준편차가 1인 정규분포는 표준정규분포라고 함
- Z변환 = (표본 – 모평균) / 모표준편차
- 정규분포의 특성을 그대로 가짐(종모양, 0을 중심으로 대칭, 면적 1)
- 용도: Z분포로 하는 검정(test)을 Z-검정(Z-test)라고 함

- **t 분포(student)**
  - 정규분포에서 파생되었기 때문에 속성이 유사하나, 데이터 개수에 따라 분포의 모양이 달라진다는 점은 다르다(정규분포는 표준편차에 의해 모양이 달라짐).
    → 수집된 표본의 관측치 수에 따라 분포의 모양이 달라짐
    → 자유도(df, n-1)라는 모수에 의해 모양이 결정됨
  - 모집단의 특성(모표준편차)을 모르고 평균을 추정 및 검정하고자 할 때 주로 사용됨(정규분포의 평균 측정)
  - 0을 중심으로 좌우 대칭되는 종 모양 분포라는 점에서 표준정규분포와 유사하지만 꼬리 부분이 더 평평하다.
    → cf. 표본 크기가 더 적으면 적을수록 분포의 꼬리는 더욱 평평해진다.

- **카이제곱분포**
  - 정규분포를 따르는 모집단에서, 크기가 n인 표본을 무작위로 반복 추출했을 때, 각 표본에 대한 표본분산들은 카이제곱 분포를 따른다.
    → 모집단 ~ N
    → 표본 랜덤 샘플링(표본 개수: n)
    → 각 표본들의 분산은 카이제곱 분포를 따름
  - 용도: 모집단의 분산을 추정할 때, 여러 집단 간의 독립성/동질성을 검정할 때(: 교재 152 페이지 카이제곱 검정 참고) 사용됨
  - 대칭되지 않기 때문에 임계치를 직접 구하기 힘듦

  cf. 표본 간 분산 차이가 없으면 검정 통계량 값이 0이 나올 수 있다.

- **F분포**
  - 분산이 같은 두 정규모집단으로부터, 각각 표본의 크기가 n1, n2인 확률표본을 반복하여 독립 추출한 후, 두 표본분산의 비율들의 표본분포를 의미한다.
    → 두 분포의 분산을 비교하는 데 활용된다.
  - ANOVA에서는 그룹 내 변동(분산), 그룹 간 변동(분산)을 통해 집단의 평균값을 비교하는 지표로 활용됨

- 회귀분석에서는 회귀모형 자체의 유의성을 검정하는 데 쓰임 (회귀계수의 유의성을 검정할 때는 t분포를 사용함)
- F분포 역시 수집된 데이터(표본의 개수)가 커지면 정규분포에 근사함

- **와이블분포**
  - 지수분포를 일반화 시켜 다양한 확률분포 형태를 모두 나타낼 수 있도록 고안됨
  - 부품의 고장까지의 시간 혹은 수명 등과 같이 신뢰성과 수명시험 문제에 적용되나, 그 이외에는 대부분 쓰이지 않음

이산확률분포

- **베르누이 분포**
  - 표본공간Sample space이 단지 두 개의 상호 배타적인 원소로 구성된 실험의 시행
    → 확률 값: p, 1-p
  - 다시 말해 확률변수 X가 단지 두 개의 값 중 하나를 취할 때(ex. 양품 혹은 불량, 성공 혹은 실패) 변수 X는 베르누이 분포를 따르는 것

- **이항분포**
  - 베르누이 시행을 여러 번 했을 때 사용되는 분포
  - 평균 = np (시행 횟수 x 기준 확률 값), 분산 = npq (시행 횟수 x 기준 확률 값, 그 외 확률 값)
  - 만일 p가 0.5에 가까워지고 시행횟수가 매우 많아지면 이항분포는 정규분포 곡선에 가까워진다(이항분포의 정규 근사)
    → (1) p가 일정한데 n이 많아질 때, (2) n이 일정한데 p가 0.5일 때 정규분포로 근사
    → np(평균)이 5 이상이고 npq(분산)가 5 이상일 때 정규분포로 근사

- **포아송 분포**
  - 일정한 단위(시간 혹은 공간)에서 발생하는 사건 발생 횟수(=X, 확률변수)에 대한 '이산확률분포'

표본 추출 Sampling

(모집단을 대표하는) 표본을 추출하는 과정/행위

- **무작위 샘플링 Random Sampling**
  - 각 단위는 선택에 대해 동일한 기회를 가짐
- **층별화 된 무작위 샘플링**
  - '단위'를 구분하여 각각 적정한 수를 무작위로 샘플링
  - 연령대별, 지역대별, 분반 별 등 특정단위
- **계통적 샘플링**
  - 일정 간격으로 샘플을 뽑아 낸다(이를 테면 일정 간격으로 제품이 생산될 때 3번 마다 제품을 하나씩 표본으로 추출)
- **서브그룹 샘플링**
  - 시간 단위로 순사대로 (ex. 2시간 간격으로 5개씩 뽑아서 표본으로 삼음)

통계량의 확률분포

표본 통계량에는 3가지가 있다: 표본평균, 표본분산(표본 표준편차), 표본비율

표본평균

- **정규분포**
  - 모수의 특징(모표준편차)을 알 때 모평균 추정
- **t분포**
  - 모수의 특징(모표준편차)을 모를 때 모평균을 추정

표본분산

- **카이 제곱 분포**
  - 1개 집단의 모분산 추정
- **F 분포**
  - 2개 집단의 분산비 비교

표본비율

- **이항분포**
  - p가 일정한데 시행횟수가 많아지거나, n이 일정한데 p=0.5일 때 정규분포로 근사함

표본평균의 분포

표본평균의 분포는 표본의 크기가 충분히 클 때 정규분포를 따른다.
→ 이때 표본평균의 평균은 모평균을 따르고, 표준편차는 모표준편차/sqrt(n)가 된다. 이를 표준오차라고 말한다.

- **중심극한정리**
  - 모집단의 형태가 어떻든지 간에 표본평균의 분포는 정규분포에 근접하게 됨
  - 모집단이 정규분포면 표본평균은 표본 크기에 상관없이 정규분포를 따름
  - 모집단이 적어도 대칭형이면 표본 크기가 5~20만 되어도 표본평균은 정규분포에 가까워짐
  - 모집단이 정규분포와 거리가 먼 분포 특성을 보이더라도, 만일 관측치의 개수가 충분히 크다면(30개 이상이라면) 표본평균은 정규분포를 따름
  - 추정과 검정을 표본평균을 바탕으로 실시하기 때문에 표본평균이 정규성을 가지는 것이 중요하다.
  - 이때 표준오차는 원래 값으로부터 벌어지는 오차를 의미하는데, 모분산보다 데이터 개수만큼 작아진다. n이 커지면 0에 근사하기 때문에, 편차가 거의 없어진다.

cf. 표본데이터의 확률분포를 활용하여 모집단의 특성에 대해 통계적으로 추정을 하는 것.

통계적 추정과 검정

통계적 추론과 검정

기술통계: 자료를 수집해서 정리, 표현, 요약, 해석하고 자료의 특성을 규명하는 기법

추론통계: 표본에 대해 그 모집단의 어떤 특성(평균이나 분산 등)에 대해 결론을 '추론'하는 절차와 기법

추론통계

통계적 추정

표본의 성격을 나타내는 통계량을 기초로 하여 모수를 추정하는 통계적 분석 방법.

- **점 추정**
  - 추정하고자 하는 모수를 표본 데이터를 활용해 하나의 수치로 추정 (ex μ = 46일 것이다)

  cf. 모분산에 대한 점 추정량은 편차 제곱의 합을 n-1로 나누어준다는 점이 특징이다.
  추정량은 절차/과정이고, 추정치는 결과적으로 산출된 수치를 의미한다.

- **구간 추정**
  - 추정하고자 하는 모수가 존재할 것으로 예상되는 구간을 정해 추정 (ex 신뢰수준 95%로 추정할 때, 모수는 -1.96과 1.96 사이에 있을 것이다)
  - 모평균 추정
    1) 모표준편차를 알 때: 정규분포 사용
    2) 모표준편차를 모를 때: t분포 사용
    (모집단이 정규분포일 경우 항상 성립, 정규분포가 아니더라도 표본의 크기가 충분히 크다면 중심극한정리에 의해 성립)

    모분산 추정
    카이제곱 분포 사용 (모집단이 정규분포일 경우 항상 성립)

    모비율 추정
    정규분포 사용 (원래는 이항분포 사용, 그러나 표본크기가 크면 근사적으로 성립)

  cf. 신뢰수준: 추정하고자 하는 모수가 신뢰구간(점 추정값 ± 한계오차) 내에 포함될 확률
  모평균 신뢰수준: 95%, 99%

  cf. 대칭인 분포에서는 신뢰수준을 구하기 쉽지만, 비대칭인 카이제곱이나 F통계량은 임계치 값을 직접 찾아내기 어렵다.

가설 검정

모수에 대해 특정 가설을 세우고 -> 표본을 바탕으로 통계량을 계산 -> 이를 기초로 (모수에 대한) 가설의 진위를 판단

cf. 모수: 모집단의 기술적 척도, 통계량: 표본의 기술적 척도

표본으로 구해지는 결론과 추정치들은 항상 옳은 것이 아니기에, 통계적 추론의 신뢰 척도로 '신뢰수준'과 '유의수준'을 활용함

- **정규성 검정**
  - 확률분포가 정규분포를 따르는지 아닌지 확인하는 검정이다. 회귀 분석 등의 기본 전제가 되기 때문에 중요하다.
  - 비정규성을 가지는 데이터:
    1) 이상요인 파악 가능
    2) 개선의 방향 제공
    3) 따라서 비정규성의 유형과 원인 파악을 할 수 있어야 함
  - 귀무가설 H0: 모집단은 정규분포를 따른다
    대립가설 H1: 모집단은 정규분포를 따르지 않는다
    → p-value가 높아서 H0을 기각하지 못해야 한다. (등분산성 검정도 마찬가지)
  - 보통은 샤피로-윌크 검정을 통해서 실시한다.

모평균의 신뢰구간 추정

- **모표준편차를 아는 경우**
  - **Z 표준정규분포**를 사용한다.
  - 신뢰도 95% 신뢰구간: 표본평균 ± 1.96 * 표준오차
    → 표본평균을 이용해 구한 범위 내에 모평균이 있을 확률 95%
    → -1.96 ≤ Z ≤ 1.96을 모평균에 대해 정리하여 구할 수 있음
    → Z(α/2 = 0.025) = 1.96임을 이용, 신뢰도 99%인 경우 Z(α/2) = 2.58
- **모표준편차를 모르는 경우**
  - **t 분포**를 사용한다. (표본으로부터 표준편차 s를 추정하여 신뢰구간을 구함
  - t(α/2, n-1)을 사용하여 구한다.

모분산 신뢰구간 추정

- 모분산에 대해 100(1 – α)% 신뢰구간 구하기
  - 비대칭분포이기 때문에 확률변수의 값을 α/2를 이용해서 구할 수 없다.
    → α/2과 1- α/2를 이용하여 구해야 한다. (자유도 n-1)

모비율p 신뢰구간 추정

- 모집단 특성이 비율에 의해 주어지는 경우(실업률, 당첨률 등) 모비율을 추정한다.
- n회 독립시행에서의 확률변수 X는 이항분포를 따른다고 할 때, 이때 비율 p의 표본분포는 정규분포를 따른다.
- 시행횟수가 충분히 많다(30번이 넘는다)는 가정 하에 정규분포로 근사하고, 표준정규분포를 이용하여 신뢰구간을 추정한다.
- 표본비율 ± Z(α/2) * sqrt{표본비율 * (1-표본비율) / n}을 통해 모비율 신뢰구간을 직접 구함

통계적 가설검정

절차

- 가설 수립: 귀무가설, 대립가설을 수립한다.
- 유의수준 결정 (보통 5%) ⇔ 신뢰수준 (95%)
- 문제 상황에 맞는 적절한 검정통계량 결정(Z, t, F, 카이스퀘어 등)
- 데이터로부터 검정통계량 계산 및 p-value 계산
- p-value < α면 H0 기각
  p-value > α면 H0 채택

용어

- **귀무가설 H0**: 기존 사실에 대한 가설, 우리가 익히 알고 있는 가설, 검정의 대상임.
  - 검정통계량은 귀무가설의 분포로부터 결정되어야 함
- **대립가설 H1**: 새롭게 확인하고자 하는 사실에 대한 가설, 주장하는 가설.
  - p-value가 유의수준보다 작을 경우, 즉 검정통계량이 귀무가설로부터 나왔다고 보기 어려운 경우 대립가설을 채택

- **유의수준**: 귀무가설이 맞는데 귀무가설을 기각하는 (= 귀무가설을 기각하는 결정이 잘못될 수 있는) 최대 가능성(확률) 제1종오류
- **임계값**: 정해진 유의수준에서 귀무가설의 채택과 기각에 관련된 의사결정을 할 때, 그 기준(귀무가설 기각의 기준)이 되는 통계량
    - 임계값을 중심으로 기각영역과 채택영역이 결정됨
- **p-value**: 귀무가설이 참이라는 가정 하에, 표본 데이터가 귀무가설을 지지하는 확률
    - 대립가설 채택을 위해서는, p-value(귀무가설을 지지하는 확률)가 유의수준(귀무가설이 맞는데 기각할 확률)보다 작아야 함 → 쉽게 생각하면 귀무가설 지지율이 최소화 되어야 함

cf. "귀무가설을 기각할 수 없다"는 귀무가설이 '옳다'는 의미가 아님. 귀무가설을 기각할 '확실한 증거가 없다'는 의미로 해석해야 함. 통계적으로 귀무가설을 기각할 수 없어도, 귀무가설은 거짓일 수도 있음 (참일 확률이 신뢰수준일 뿐)

가설검정의 오류

- **제1종 오류 α**
    - 생산자 위험 (귀무가설을 기각함으로서 새로운 투자를 해야 하기 때문)
    - 귀무가설을 채택해야 했는데, 기각하는 위험도를 의미.
    - 보통 5%로 결정하며, 귀무가설 기각 시 위험도가 높으면 1%로 설정
- **제2종 오류 β**
    - 소비자 위험(소비자가 감당해야 하는 리스크)
    - 귀무가설을 기각해야 했는데, 채택하는 위험도를 의미.
    - 보통 10% 설정
    - 1종오류와 2종오류는 trade off 관계
    - (1 - β) = 검정역 (cf. 적절한 표본 개수를 검정역을 바탕으로 정함)

평균 검정 t-test

- 표본평균과 표본 표준편차를 이용해 모집단의 평균 및 평균의 차이를 검정함
    - 2 sample의 경우 평균의 차이가 표본 오차에 의한 것인지(귀무가설), 아니면 두 모집단의 속성에 의한 것인지(대립가설) 밝히는 것을 의미
- t-검정의 강건성: 모집단이 극단적으로 비정규성을 띠지 않는 이상, t-검정과 신뢰구간 추정치

는 여전히 타당함

- 정규성 검정과 등분산성 검정(Levene)이 선행되어야 함.
  - 독립성이 만족되지 않는 경우 → Paired t-test (같은 대상이니까 독립X)
  - 정규성이 만족되지 않는 경우 → Mann-Whitney test 시행
  - 등분산성이 만족되지 않는 경우(분산이 서로 다름) → 자유도 수정한 2 sample t-test
  - 독립성, 정규성, 등분산성이 모두 만족되면 자유도 수정 않고 2 sample t-test
- **귀무가설 H0: 두 집단의 평균은 같다**
  대립가설 H1: 두 집단의 평균은 같지 않다
  유의수준: 보통 0.05
  검정통계량 t값 계산
  검정 결과 판단: p-value와 α값 비교하여 귀무가설 채택/기각 여부 결정 (t값의 통계적 유의성 검정)
- t-test 종류
  - **1-sample t test**
    : 단일 집단의 평균이 특정 값과 같은지 비교
    (ex. 대한민국 고3의 평균 수면시간이 정말 6시간이 맞을까?)
    - ✓ 모표준편차를 모를 때 사용한다 (알면 t-test가 아니라 z-test 시행한다)
    - ✓ **귀무가설: 평균이 (기존과) 다르지 않다**
      대립가설: 평균이 (기존과) 다르다
  - **2-sample t test**
    : 두 집단 간 평균이 같은지 비교
    (ex. A반과 B반 사람들의 키 평균은 같은가?)
    - ✓ **정규성과 등분산성 검정을 선행적으로 진행**함
      → 세 가지 가정을 만족해야 하기 때문
      1) 각 데이터는 독립적이고 랜덤으로 샘플링 되었다
      2) 두 그룹은 각각 정규분포를 따른다
      3) 각 그룹의 분산은 같다
    - ✓ 두 집단이 정규성을 따르지 않는다 -> 등분산성에서 Levene test
      두 집단이 정규성을 따른다 -> 등분산성에서 F test
      두 집단의 분산이 같다(귀무가설) -> Student t-test
      두 집단의 분산이 다르다(대립가설) -> Welch's t-test

✓ **귀무가설 H0: 두 집단의 평균은 동일**
대립가설 H1: 두 집단의 평균은 다름

- **Paired t-test**
  : 쌍을 이루는 두 집단 간 평균이 같은지 비교(앞/뒤, 전/후)
  (ex. 포스코 교육을 들어온 뒤 사람들의 몸무게 전후 평균은 같은가? = 포스코 교육을 들은 이후 사람들의 몸무게에 변화가 있는가?)

  ✓ **가정**
  1) 종속변수 y가 연속형 변수(등간척도, 비율척도)
  2) 변수는 서로 독립적
  3) 종속변수 y는 대략적으로 정규분포를 따름
  4) 종속변수 y는 이상치를 포함하지 않음

  ✓ **귀무가설 H0: 전후 평균 차이(Before – After)가 없다**
  대립가설 H1: 전후 평균 차이(Before - After)가 있다
  → 평균 차이가 양수이면 Before(이전)의 평균 값이 더 큼
  → 평균 차이가 음수이면 After(이후)의 평균 값이 더 큼 (개선효과, 점수 상승 등)

평균 검정 ANOVA

- **집단들 간의 평균 차이를 검정**
  → 총 변동을 집단 간 변동, 집단 내 변동으로 분해함
  → 이 두 변동의 비가 통계적으로 유의한지를 검정함 (분산 비율: F비)

- **귀무가설 H0: 모든 집단의 평균은 같다**
  대립가설 H1: 적어도 하나의 집단은 평균이 다르다
  (평균 차이가 나는 집단이 적어도 하나 이상 있다)

- 집단 간 평균 차가 없으려면 (= **귀무가설 채택**하려면)
  - **그룹 간 차이가 적고, 그룹 내 차이가 커야 함**

- 집단 간 평균 차가 있으려면 (= **귀무가설 기각**하려면)
  - **그룹 간 차이가 크고, 그룹 내 차이가 작아야 함**

cf. ANOVA Table 채울 수 있어야 함
cf. 귀무가설을 기각하더라도, 어떤 집단에 차이가 있는지를 모르기 때문에 boxplot을 그린다
cf. 회귀분석에서는 회귀모형의 적합도를 판별하는데에도 분산분석을 이용한다.

비율 검정

- **1 Proportion-test**
  - 한 집단의 비율이 특정 비율과 같은지를 검정
    (ex. 과거의 불량률이 10%였다. 지금은 9%인데, 과연 통계적으로 유의미한 차이가 없는가?/있는가)
- **2 Proportion-test**
  - 두 집단의 비율이 같은 지를 검정
    (ex. 동일 제품을 생산하는 공장에서, 불량률은 통계적으로 유의미한 차이가 없는가?/있는가?)

카이제곱 검정

- **동일성 검정(비율 검정)**
  - 두 가지 이상의 범주 간에 어떤 비율이 상호 동일한 비율로 나타나는가를 검정
    (ex. 포스코 교육 기수 별 이탈율에 차이가 있는가?)
- **독립성 검정**
  - 두 가지 이상의 범주 간에 상호 관련성이 있는가?
    (ex. 자동차 브랜드와 거주지 사이에 상호관계가 있는가?)
- **적합도 검정**
  - 어떤 특성이나 사건이 기대치에 따라 발생했는가?
    (ex. 완두콩 수확량이 각각 n1, n2, n3, n4일 때 이는 멘델의 유전 법칙과 일치하는가?)
- 카이제곱 검정 통계량 값이 크다 = 실측치와 기대치 차이가 크다
  → 귀무가설 기각 (귀무가설: 차이가 없다)
  카이제곱 검정 통계량 값이 작다 = 실측치와 기대치 차이가 작다
  → 귀무가설 채택 (귀무가설: 차이가 없다)

상관/회귀분석

상관분석

- 두 수량형 변수 간의 선형적 관계의 강도(얼마나 선형성을 가지는가), 방향(양/음)을 분석

선형관계의 척도

- **공분산 covariance**
  - 둘 이상의 변량이 연관성을 가지며 분포하는 모양을 전체적으로 나타냄
  - 척도 단위에 영향을 많이 받음.
- **상관계수**
  - 두 변수 간 선형 관계의 정도와 방향을 수치로 표시함
  - 표준화 하였기에 변수 척도 단위에 영향을 받지 않음
  - -1에서 1 사이의 값을 가짐(-1에 가까울수록 음, 1에 가까울수록 양의 상관관계)
  - 두 변수 간 연관된 정도를 나타날 뿐 인과관계를 설명하는 것은 아님!!
- **산점도**
  - x, y 두 변수 간 대략적 관계를 눈으로 쉽게 알아볼 수 있음(시각화 plot)
  - 그러나 두 변수 간 관계를 정확히 파악하지는 못함(따라서 수치로 정량화하여 표현한 상관계수를 확인하는 것임)
  - 착안사항
    1) 점들이 산재된 모양으로부터 두 변수 사이의 관계 여부 확인
    2) 직선관계인가 곡선관계인가 확인
    3) 이상 데이터 여부 확인
    4) 점들이 몇 개의 집단으로 층별 되는지 확인

회귀분석

- 독립변수x가 종속변수y에 미치는 영향력 크기를 측정
  → 독립변수x의 일정한 값에 대응되는 종속변수y의 값을 예측하기 위한 통계 분석 방법

단순선형회귀

- 목표변수 y 1개, 설명변수 x 1개
- 회귀계수(=기울기)는 최소 자승법을 활용해 추정함

- 기울기는 x가 y에 미치는 영향력의 정도(크기)를 나타냄(0에 가까울수록 영향력X)

- **회귀모형의 적합도 판정**
  (아래의 순서대로 분석이 시행됨)

  - **ANOVA 분산분석으로 회귀모형 적합성 검증**

    ✓ SSR에서 df는 변수의 개수, SST의 df는 n-1, SSE의 df는 (n-1) – SSR의 df임. 이 값을 직접 구할 수 있어야 함.

    ✓ MSR은 SSR을 df로 나눠준 값, MSE는 SSE를 df로 나눠준 값

    ✓ MSR과 MSE를 나누어서 F비를 구하고, 이를 검정하여 p-value 값이 유의수준보다 낮으면 회귀모형이 적합하다고 판단함.

    ✓ 귀무가설 H0: 회귀계수 = 0 (설명력 없음)
    대립가설 H1: 회귀계수 ≠ 0 (설명력 존재)

    cf. 다중선형회귀모형의 ANOVA 귀무/대립가설
    귀무가설 H0: 모든 회귀계수 = 0
    대립가설 H1: 하나의 회귀계수라도 0이 아님

  - **결정계수(R square) 확인**

    ✓ SST(y의 총변동) = SSE(잔차의 변동) + SSR(회귀식으로 설명 가능한 변동)
    cf. SSE는 회귀식으로 설명 불가능한 변동임.

    ✓ 결정계수는 SSR/SST로 구하며, 0에서 1 사이의 값을 가짐
    → 1에 가까울수록 모형이 적합하다고 판단

    ✓ 수정된 결정계수: 다중회귀분석에서는 x가 많아질수록 결정계수 값이 높아진다는 결점이 있음(SSR이 높아지기 때문임. x가 아무리 쓸모 없는 변수라 하더라도, 분산 값은 0 이상이기에 총변동을 어느정도 설명함. 따라서 SSR은 무조건 높아짐). 이러한 한계를 보완하기 위해 수정된 결정계수를 사용

  - **t 검정을 통해서 회귀계수의 유의성 확인**

    ✓ 귀무가설 H0: 회귀계수 = 0
    대립가설 H1: 회귀계수 ≠ 0

  - **잔차항의 필요조건 확인**

    1) 잔차항이 **정규성**을 따르는가? (평균 0) → QQ plot이나 히스토그램으로 확인

    2) 잔차항이 **등분산성**을 가지는가? → 깔대기 효과가 나면 안 됨. 그래프에서 경향성 없이 찍혀야 함.

3) 잔차항이 서로 **독립**인가? → 가로축 기준으로 일정하게 모여 있어야 함. 경향성을 보이면 안 됨.

- 최종모델 선정

cf. 파이썬으로 OLS 모듈을 이용하여 구함

cf. 다향회귀

- **다중 공선성** 문제 발생(독립변수 x들 사이에 상관관계/선형관계가 있는 현상)
  → 다중 공선성이 있으면 회귀 계수의 분산이 매우 커짐. 결정계수 및 F값의 신뢰도가 낮아짐.
- **확인 지표: VIF** (보통 **5~10 넘어가면 다중 공선성이 있다**고 판단)
  → 도메인 지식을 근거로 판단했을 때, VIF가 10 이상이어도 중요한 변수라면 삭제하지 않음