통계분석 마스터 클래스 스터디 메모

📌 강의 메모 중 주요 내용과 직접 찾아본 추가 지식을 대강 메모함 (추후 정리 필요)

Section 1

- 통계학을 통해 어떻게 다음을 해결할 것인가?
  1. 어떻게 결과를 해석할 것인가?
  2. 어떻게 그 결과를 실무에 적용할 것인가?
- 통계 기반으로 데이터 분석 공부하기
  1. 이해도 - 피어슨 통계학의 전 파트를 다루면서 이해
  2. 숙련도 - 실제 실습을 통해 반복적인 이론 꺼내기 연습

▼ 수업에서 제시하는 큰 강의 흐름

1. 의사결정 Framework
   - PPDAC 모형 - 비즈니스 문제를 어떻게 도출할지와 통계를 가지고 분석해서 실제 의사결정까지 결정하기 위한 모형
2. 통계학 (Pearson 통계)
   - Pearson 통계학은 주류 통계학으로 흔히 사용됨
     1. 기술통계 - 데이터를 요약해서 인사이트를 뽑아내는 기술 (데이터분석)
        - 값을 잘 요약하기
        - 요약 결과를 시각화해서 해석하기
     2. 확률론 - 기술통계로 도출한 표본이 어떻게 모집단을 추론하기 위한 충분한 설명력을 갖는지 알아내는 이론
        - 확률론의 큰 흐름
          1. 확률론의 원칙
          2. 확률변수에 대한 이해
          3. 확률분포에 대한 이해
        - 수학적 방법을 Data Bootstrap 방법으로 시뮬레이션하여 설명
     3. 추론통계 - 기술통계와 확률론을 이용해서 추론 결과가 맞는지 검증하는 것
        - 기본지식
          - 카이제곱검정 - 문자vs문자 관련성 검증
          - T검정 - 문자vs숫자 관련성 검증 (2그룹)
          - ANOVA (3그룹 이상)
        - 회귀분석 - 단일/다중 회귀분석
        - 로지스틱 회귀분석
          - ML을 통해 예측하는 방법을 통해 이해

Section 2

- 최적의 의사결정?
  - 직관 + 데이터 = 더 나은 결정
  - 경영 = 의사결정
    - 결국 비즈니스에서 의사결정하려는 주제가 무엇인가?
    - Business Question으로도 부름

1. **직관적 의사결정 (경험에 따른 직관)**
   - 고객의 욕구, 경쟁자 전략, 기술변화 등에 대해 실제 경영상황에서는 완벽하게 정보를 획득하고 대안을 도출하기 힘듬
   - 제한된 합리성 하에서의 의사결정을 내리는 방법
     - 사이먼의 관리적 의사결정 모형

    - 오랜 현장경험과 노하우가 있는 경영자는 부족한 정보 하에서도 가능한 대안을 파악하고 핵심을 찾아내는 직관적 능력을 보유
    - 즉, 경험과 직관의 중요성을 강조하는 사례

2. **데이터 기반 의사결정**
    - 왜 데이터 기반 의사결정이 더 뜨고 있는가?
        - 오프라인 시장 규모의 하락과 디지털 시장의 상승으로 인해 데이터 기반 의사결정이 더 활성화되는 중
            - 전통 비즈니스 시장의 특성
                - 수십년의 업력을 가지고, 직관 기반 의사결정을 진행할 수 있는 충분한 경험이 확보됨
            - 디지털 시장 특성
                - 최신 기술이나 트렌드에 맞게 발생하는 새로운 시장에서, 의사결정을 위한 충분한 업력이 부족한 시장
    - 직관적 의사결정과 대칭점에 있는 의사결정 모형
        - 무작위 표본을 기반으로 모집단을 모형으로 나타낸 뒤 모형을 활용해서 통계적 추측을 진행
        - 확률 = 불확실성의 정도를 측정한 값
            - 사건 발생의 개연성을 측정하는 수단으로, 합리적 의사결정을 내리기 위해 알아야할 정보의 차이를 수량화한 도구
        - 즉, 이러한 확률의 도입을 통해 확실성이나 부족한 지식에 대한 대체재 역할을 수행함
    - **데이터 사이언티스트 필요 역량**
        1. 분석도구 매뉴얼 활용 역량
            - Python, Excel, SPSS, Tensorflow, ...
        2. 통계 지식
            - 피어슨 통계 (기술통계, 확률론, 추론통계)
            - 베이즈 통계 (조건부 확률, 축차합리성)
        3. 분석 모델
            - ARM, AARRR 등, 예측, 군집화, 연관성 규칙 발견, 분류 앙상블
        4. 비즈니스 도메인 지식 & 경험
            - 호텔, IT, 광고 등 DS 적용 대상 실무 분야
    - **통계적 의사결정 모형**
        - **PPDAC 모형**
            - Problem (문제) - 문제를 이해하고 정의하기
                - 이 문제에 답하려면 어떻게 시작할까?
            - Plan (계획)
                - 무엇을 어떻게 측정할 것인가?
                - 연구 설계와 데이터 수집은 어떻게?
            - Data (데이터)
                - 데이터를 수집/관리/처리하기
            - Analysis (분석)
                - 데이터 요약하기
                - 표/그래프 만들기
                - 데이터 분류하기
                - 패턴을 찾아내기
                - 가설을 세우기
            - Conclusion (결론)
                - 분석 결과를 해석하기
                - 결론을 내리기
                - 이를 잘 전달하기

Section 3

- **데이터의 유형과 측정 기준**
    - 대상의 속성을 숫자 또는 문자로 표현한 것
    - 데이터는 문자 또는 숫자이다.

    - 데이터는 범주형(문자) 또는 연속형(숫자)으로 분류한다
        - 범주형 (질적 변수)
            - 명목척도 - 남자는 1, 여자는 2로 표현
            - 서열척도 - 1등, 2등, 3등
        - 연속형 (양적 변수)
            - 등간척도 - 질병의 통증을 1-10으로 표현
            - 비율척도 - 거리, 넓이, 무게, 금액 등

- **정형 데이터**
    - 통계적 분석의 대상이 되는 데이터
    - 정해진 규칙에 맞게 데이터를 설계 및 보유하는 형태
    - 데이터베이스의 표(테이블) 형태로 관리하는 것이 일반적
        - 열(Field), 행(Record)로 구성된 테이블 데이터
        - 분석 도구들의 분석 기능을 문제없이 사용하기 위해 데이터는 테이블 형태여야함
            - 하나의 열에 하나의 속성을 갖도록
            - 필드명은 단일 행으로만 구성되도록
            - 레코드는 위에서 아래로 구성되어야함
    - 데이터 분석 = 데이터 요약 = 열 데이터를 요약하는 것
        - 열 - 데이터에서 요약하려는 대상
        - 행 - 모집단을 이해하기 위한 표본

- **DIKW 피라미드**
    - 데이터 기반 비즈니스 전략 수립 방법
    - 데이터(Data) → 정보(Information) → 지식(Knowledge) → 지혜(Wisdom) 순으로 높아지는 피라미드 형태
    - DIKW 피라미드의 예시
        1. 데이터
            - 21:40분에 40대 초반 남성이 계속 온라인 몰에서 2개 상품 페이지를 터치함
            - 그 2개 상품은 소니 헤드폰과 애플 헤드폰
        2. 정보 (or Context)
            - 위의 맥락으로부터, 특정 시간대+40대남자+헤드폰의 구매를 고려중
        3. 지식 (or Meaning)
            - 결국, 제품을 구매할 것인지 고민하고 있다는 것을 알게 됨
        4. 지혜 (or Insight)
            - 구매 고민을 해결하기 위해 소니의 중저가 헤드폰 신상을 추천하면 즉시 구매할 가능성이 높아질 것이라 예상함
            - 이에 따른 자동화 추천을 진행

Section 4

- 데이터 확보 - 1차 자료와 2차 자료
    - 1차자료 - 직접 관찰 및 수집하거나 사람들에게 조사하여 얻는 자료
        - 서베이 (survey)
        - 직접 측정하는 실험 관찰 데이터
    - 2차자료 - 과거에 다른 목적이나 용도로 수집되었던 조직화된 정보에서 목적에 적합한 데이터를 활용하는 자료
        - 데이터베이스에서 추출
        - 크롤링 (crawling)

Section 5

- 기술 통계의 필요성
    - 데이터 요약 - 기술 통계

        - 대표적인 축약 방법
            - (시각화) 그래프로 만들어 그 특징이나 패턴을 파악
            - (값으로 요약) 숫자 하나로 데이터 특징을 대표하도록 함 (대표값을 '통계량'이라고 함)
- '통계량' - 데이터를 값으로 요약하기
    - 평균 - 데이터의 무게중심에 해당하는 값
    - 중앙값 - 크기 순서대로 나열해서 중간 지점에 해당하는 값
    - 편차(관찰값 - 평균값) - 평균값으로부터 어느 정도 큰지, 작은지를 나타냄
    - 분산 - 데이터가 평균으로부터 얼마나 퍼져있는지 평가
    - 표준편차 - '평균값'이 데이터의 분포를 대표하는 수치지만, 표준편차는 그 대푯값을 기점으로 해서 데이터가 대략 어느 정도 멀리까지 위치해 있는지를 나타내는 통계량

Section 6

- **평균과 표준편차의 해석**
    - **표준편차는 확률적으로 일반적/특별함을 구별하는 기준이다**
        - 표준편차 1배 범위 안의 데이터 = 일반적인 데이터라고 할 수 있음
            - 정규분포와 같은 대칭분포 가정 시, 68.27% 확률 이내
        - 표준편차 2배 범위 밖의 데이터 = 특별한 데이터라고 할 수 있음
            - 정규분포와 같은 대칭분포 가정 시, 95.45% 밖의 확률

▼ 버스 도착 시각으로 알아보는 통계량 - 평균, 편차, 분산, 표준편차

- 목적: 어떻게 하면 버스를 놓치지 않고 최적 시간에 버스정류장에 도착할 수 있을까?
    - → 버스를 놓치지 않고 탈 수 있는 최적시간은?
- 오전 7:30의 버스 5회 관찰 데이터
    - 평균값 = 31 (단위: 분)

관찰회차	도착시간 (분)
1	32
2	27
3	29
4	34
5	33

- 평균과 실제의 차이를 가지고 분석하기
    - 편차제곱의 평균 (분산) = 6.8 (단위: 분의 제곱)
    - 분산의 제곱근 (표준편차) = 2.607... (단위: 분)

관찰회차	도착시간 (분)	평균값	편차	편차제곱
1	32	31	1	1
2	27	31	-4	16
3	29	31	-2	4
4	34	31	3	9
5	33	31	2	4

- 대칭적인 모양의 이 버스 관찰 데이터가 어떤 분포라고 가정한다면
    - 1표준편차 이내 - 약 70% 이내 → 28.4분 ~ 33.6분 이내에 70%확률로 도착
        - 놓칠 확률: 30% 확률 중 버스도착시간보다 빨리 오면 되는문제이므로 숫자가 작은 쪽은 고려할 필요 없는 단식 검정 문제
        - 즉, 15%확률로 놓칠 수 있고, 100번 도착할 때 15회는 놓칠 수 있음
    - 2표준편차 이내 - 약 95% 이내 → 25.8분 ~ 36.2분 이내에 95% 도착
        - 놓칠 확률: 위와 마찬가지로 단식 검정이므로 5%가 아닌 2.5%가 놓칠 확률
        - 즉, 100번 중 2.5회는 놓칠 수 있음
- 이를 토대로, 내가 감당할 수 있는 리스크 수준에 따라서 의사결정하면 됨
    - 내가 감당할 수 있는 리스크는 N%이내임을 고려해서 선택하면 됨

▼ 비즈니스 문제에서의 표준편차 활용 -VIP유저의 결제금액기준 정하기

- A스토어의 최우수 고객(상위 0.5% 결제자)를 특별관리하고자 함

        - 단골 고객 30,000명의 데이터를 분석한 결과, 고객 당 평균 결제액은 500,000원이며, 표준편차는 250,000원이었다.
        - 적절한 최우수 고객 기준 결제액은 얼마인가?
            - 상위 0.5% 결제자를 골라야하므로, 분포의 99% 인 3표준편차 밖의 유저를 선택
            - 분포에서 3표준편차 밖인 하위 0.5%와 상위 0.5% 중 상위 0.5%만 선택
            - 즉, 이 금액기준은 평균 + 3표준편차 = 500,000원 + 3 * (250,000원) = 1,250,000원
            - 결과적으로, 고객 당 최우수 고객 기준 결제액은 125만원 이상임
    - 시사점: 주요 통계량에서 평균과 표준편차만 잘 활용해도 상당히 의미있는 의사결정을 내려줄 수 있음

- **상관관계와 상관계수**
    - 상관관계 = 한 변수가 증감할 때 다른 변수가 얼마나 증감하는지를 나타내는 관계
    - 상관계수 = 관계를 -1 ~ 1로 수치화한 것
        - 두 변수에 대한 공분산에 두 변수에 대한표준편차곱을 나눔
        - 피어슨 상관계수
            - 두 변수 모두 연속형 자료, 정규성을 따른다는 가정 필요
        - 스피어만 순위 상관계수
            - 비모수적 방법, 순서형 변수에 대해 사용 OR 정규성을 벗어나는 경우 사용
    - NOTE:
        - 모수적 방법: 모수를 특정 분포로 가정하여 접근
        - 비모수적 방법: 모수를 특정 분포로 가정하지 않고 접근
            - 정규성을 만족하지 않거나 표본의 개수가 10개 미만
- **Heatmap 시각화**
    - 두 양적 변수의 상관관계, 상관계수를 시각화하는데 유용
    - **다중공선성**에 대한 정보도 제공함
        - 다중공선성 (Multicollinearity)
            - 통계학의 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제
            - 흔히, 회귀분석에서 사용된 모형의 일부 설명변수가 다른 설명변수와 상관 정도가 높아서 두 개 이상의 변수 중 다른 하나의 영향력을 완벽히 통제할 수 없게 되는 문제
            - 즉, 다중공선성이 생기면 다중공선성애 해당하는 변수들 각각의 설명력이 약해짐
            - 이는 변수들의 표준오차 (Standard Error)의 증가로 나타남

Section 8

- 모집단 추정
    - 모집단: 분석의 목표가 되는 집단으로, 실제 우리가 알고자 하는 집단
        - 관심의 대상이 되는 모든 데이터 집합
    - 표본: 모집단의 일부를 선택한 집단
        - 모집단 중 조사대상으로 채택된 일부 데이터 집합
    - 표본의 데이터를 가지고 인사이트를 도출한 다음, 그 인사이트를 토대로 모집단도 그러할 것이라고 추정하는 것
- 모수와 추정량
    - 모수(Parameter) : 모집단의 특성을 나타내는 값
        - 모평균, 모표준편차 등
    - 추정량(Estimator) : 표본으로부터 모수를 추정하기 위한 값
        - 표본평균, 표본표준편차 등
- 점 추정과 구간 추정
    - 추정 (Estimation)
        - 모수를 맞치고자 하는 것 (근삿값 구하기)
        - **점 추정** - 모수를 1개의 값으로 추론
            - 성인남성 표본평균이 174.2cm → 모집단도 그러할 것이다
            - 점 추정만으로는 추정값을 얼마나 신뢰할 수 있는지 알 수 없기 때문에 구간 추정이 보다 실용적임
        - **구간 추정** - 일정한 구간 내의 어느정도 정확도로 존재

        - 성인남성 표본평균이 172.2cm ~ 176.2cm
- 구간 추정의 핵심 개념
    - 실무적으로 많이 활용하는 추정 방법
        - 모수를 포함하는 구간의 확률을 지정해야함
            - 일반적으로 임의로 95%, 99%로 지정하며, 이 확률을 신뢰계수라 함
        - 신뢰계수를 95%로 정하고, 확률 95%로 모수를 포함하는 구간 (신뢰하한 ≤ 모수 ≤ 신뢰상한)을 도출해냄
            - 이 구간을 모수에 대한 신뢰계수 95%인 신뢰구간이라 지칭함
        - 신뢰구간의 신뢰하한과 상한은 모두 표본의 통계량을 통해서만 도출해냄
            - 이 의미는 추정하고자 하는 모수가 알맞은 추정구간 내에 존재한다고 믿을 수 있는 정도
        - 구간 추정에서는 신뢰수준, 신뢰구간을 같이 제시함
- 신뢰계수가 크면서 신뢰구간의 폭이 좁은 것을 이상적으로 생각함
    - 보통 통계량을 구하는 표본 데이터가 늘어나면 신뢰계수를 낮추지 않아도 신뢰구간의 폭이 대부분 좁아지게 됨
- 신뢰수준
    - 95% 신뢰수준의 의미
        - 신뢰구간을 100번 측정한다면 그 중 95개는 모수를 포함하고 있음을 나타냄
        - 구간추정
            - 점추정 값 + 오차범위 ≤ 표본평균 ≤ 점추정 값 + 오차범위

Section 9

- 확률의 정의
    - 표본공간: 얻을 수 있는 모든 가능한 결과의 전체 집합
    - 사건: 어떤 조건을 만족시키는 결과들의 집합 (표본공간의 부분집합)
    - 확률 (A) = 원하는 결과의 경우의 수 / 모든 가능한 결과의 경우의 수
        - 즉, 사건 A 원소의 개수 / 표본공간 S의 원소의 개수
    - 동전 던지기에 대한 예시 1)
        - 표본공간 S = { (앞,앞), (앞,뒤), (뒤,앞), (뒤, 뒤) }
        - 앞면이 2번 나오는 사건 A = { (앞, 앞) }
        - 확률 = 1 / 4
- 확률변수
    - 확률에 따라 변하는 값을 의미함 → “관계”
        - 특정 확률로 발생하는 각각의 결과를 수치적 값으로 표현하는 변수
    - 동전 던지기에 대한 예시 2)
        - 표본공간 S = { (앞,앞), (앞,뒤), (뒤,앞), (뒤, 뒤) }
        - 확률변수 X가 동전을 던져서 앞면이 나온 횟수라고 할 때
        - 횟수 R = { (앞,앞) → 2, (뒤,뒤) → 0, (앞,뒤) → 1, (뒤,앞) → 1 }
- 확률분포
    - 모집단을 수학적으로 표현한 것을 의미함
    - 동전 던지기에서 앞면이 나온 횟수에 대해 표본공간 S → 실수 R 로 대응하는 확률변수 X에 대하여 확률을 계산한 모든 결과를 나타낸 것을 X의 확률분포라고 함
- 확률변수와 확률분포
    - 확률변수는 확률에 따라 발생한 결과를 수치적 값으로 표현한 “관계”
        - ex) 앞면이면 ‘1’, 뒷면이면 ‘0’으로 대응
    - 확률분포는 확률변수가 갖는 모든 경우의 확률을 계산한 것
        - 전체 시행에서 앞면이 나타난 횟수에 대한 확률분포
    - 특정 확률변수의 확률분포를 알면 특정 사건이 일어날 확률을 계산(예측)할 수 있음
- 확률 법칙에 대한 복습
    1. 0 ≤ 확률 ≤ 1
    2. 일어날 확률 = 1 - 일어나지 않을 확률
    3. 덧셈법칙 - 동시에 일어날 수 없는 배반사건의 전체 확률은 더해서 구한다.

4. 곱셈법칙 일련의 (한 사건이 다른 사건에 영향을 미치지 않음을 뜻하는) 독립사건들이 일어날 전체 확률은 곱해서 구한다.

- **데이터 부트스트랩**
  - 복원추출을 반복하는 방식으로 추정값의 변동성에 대한 아이디어를 얻는 방법
  - 자기 부츠 손잡이를 잡아당겨 스스로를 들어올린다는 것과 같다고 하여 Bootstrap 이란 이름이 붙음
  - 모집단 분포와 관련해서 어떤 가정도 하지 않고서도 추정값의 변동성에 대해서 배울 수 있다는 의미
  - 재표본 추출을 1,000번 반복하면 평균값이 1,000개 나오는데, 그 분포에 대한 히스토그램을 보면 원래 표본의 평균 근처에 부트스트랩 추정값들이 퍼져있음을 알 수 있음
  - 이것을 표본분포(Sampling Distribution)이라고 함
  - 결론적으로, 이 부트스트랩 분포를 이용해 추정값들의 불확실성을 수치화할 수 있음
  - 부트스트랩은 어떠한 강력한 가정 없이 확률 이론을 이용하지 않고 추정값의 불확실성을 평가하는 직관적이고 컴퓨팅적인 방법
- **대수의 법칙 (Law of Large Number)**
  - '시행이 많아질수록 경험적확률은 수학적확률에 가까워진다'
  - 표본의 크기가 커짐에 따라 표본평균은 확률적으로 모집단의 실제 평균값에 수렴한다.

Section 11

- **중심극한정리** (Central Limit Theorem)
  - 표본평균들이 이루는 표본분포와 모집단 간의 관계를 증명함으로써 수집한 표본의 통계량을 이용해 모집단의 모수를 추정할 수 있는 확률적 근거를 제공함
  - **모집단 분포에 상관없이 큰 표본들의 표본평균의 분포가 정규분포로 수렴한다**
- 무작위 추출 (Random Sampling)
  - Big Question: 표본이 모집단을 대표하는가?
  - 가장 간단하고 대표적인 방법 → 단순무작위추출
  - 모집단의 특정 부분만 추출해서 편향이 생기지 않도록 하기 위해 무작위로 골고루 추출되도록 하는 것이 목적
- 적절한 표본 크기
  - 표본 크기가 중심극한정리와 대수의 법칙에 의해 중요하게 작용함
  - 모집단 크기와 신뢰수준에 따라 표본오차를 계산하여, 그에 필요한 표본크기를 계산할 수 있음
    - 신뢰수준 95% → 표본오차 5%

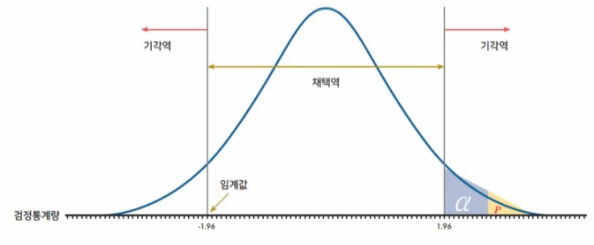
Section 12

- 인과관계
  - 독립변수와 종속변수
    - **독립변수** : 영향을 주는 변수, 가설의 원인이 되는 변수, 종속변수에 영향을 미치는 선행조건
    - **종속변수** : 가설의 결과가 되는 변수, 자극에 대한 결과나 반응을 나타내는 변수
- 인과관계와 상관관계의 이해
  - **상관관계가 있다고 해서 인과관계를 나타낸다고 할 수 없다**
    - 앱 삭제와 구매 포기가 양의 상관관계가 강하다고 나옴
    - 과연 앱 삭제 증가 때문에 구매 포기가 많아졌다고 할 수 있는가?
  - 왜냐하면, 우연한 제 3요소에 의해 두 변수에 상관관계가 나타날 수 있기 때문임
  - 인과관계를 알아보기 위해선 A/B테스트와 같은 무작위 통제 실험을 설계하여야 함
- **A/B테스트의 인과관계 조사 설계**
  - 조사의 내적 타당성과 외적 타당성을 만족하도록 설계하는 것이 중요하다.
    - 내적 타당성 - 실험 처치 이외 모든 변수는 대조군/실험군 모두 동일하도록
    - 외적 타당성 - 조사 대상을 무작위/확률적으로 추출해 대표성을 높이도록

	내적 타당성	외적 타당성
검토 대상	• 실험 결과가 정말 실험 처치(조작된 독립변인) 때문에 일어난 것이라고 볼 수 있는가? • 인과관계를 얼마나 확신할 수 있는가?	• 실험 결과를 다른 대상과 상황에게 어느정도 일반화 시킬 수 있는가?
통제 방법	• 실험 처치 이외 모든 변수는 실험군 & 대조군에 동일하도록 만든다.	• 조사 대상을 무작위/확률적으로 추출하여 대표성을 높인다.

Section 14

- 가설 검정의 원리
  - 네이만-피어슨 추론 방식
    1. 모집단에 대한 가설 수립
    2. 모집단에서 추출한 표본이 유의할 확률 계산
    3. 가설의 기각/채택
  - 네이만-피어슨 추론에 필요한 이론들
    - 귀무가설과 대립가설이 양립하지 않음
      - 귀무 참 → 대립 거짓
      - 대립 참 → 귀무 거짓
    - **귀무가설 (H0)**
      - 우리가 알고자 하는 모집단의 특성에 대한 잠정적인 주장
      - 보편적으로 알려진 기존의 사실
      - 둘 간의 관계가 없다, 차이가 없다, 영향을 주지 않는다 등
    - **대립가설 (H1)**
      - 귀무가설이 기각되면 대안으로 채택되는 가설
      - 이 분석이 의미있는 새로운 사실을 발견해냈다는 주장
      - 둘 간의 관계가 있다, 차이가 있다, 영향을 준다 등
  - **유의확률과 유의수준**
    - 유의확률 = p 로 표기
      - 귀무가설을 지지하는 힘
        - p-값은 귀무가설이 참이란 전제 하에, 표본에서 실제로 관측된 통계치와 같거나 더 극단적 통계치가 관측될 확률
        - 즉, p-value가 작을수록 귀무가설을 지지하는 정도가 약하다고 보며, 신뢰수준을 설정함에 따라 대개 0.05, 0.01보다 작은 경우 귀무가설을 기각하는 원리
    - 유의수준 = alpha 로 표기
      - 귀무가설의 기각 여부를 결정하는데 사용되는 기준이 되는 확률
        - 구간 추정에서의 신뢰계수 설정과 같은 개념
        - 95%신뢰도를 기준으로 한다면 1-0.95 = 0.05가 유의수준 값이 됨
        - 1종 오류를 범할 확률의 하용한계
      - 표본을 토대로 내린 결론이 틀렸을 때 감당할 수 있는 한계
        - 1종오류?
          - 귀무가설이 참인데 대립가설이 참이라고 잘못 판단하는 오류
          - 차이가 없는데 차이가 있다고 하는 오류
  - **유의성 검정 한장 요약**
    - 중심극한정리를 가정할 때 정규분포가 된 표본분포의 검정통계량에 대한 가설검정의 관계들을 표현



- **상황에 따른 가설검정 분석 로드맵**
  - 가장 기본이 되는 분석법에 대한 로드맵 정리
  - 유의확률 p를 구해서 가설을 기각/채택하는 원리는 동일함
  - 단, 독립변수와 종속변수의 상황에 따라 다른 방식의 분석법이 존재함

		독립변수	
		질적변수	양적변수
종속변수	양적변수	**T-검정** 1개 또는 2개 그룹의 평균값을 검정	단순선형회귀분석 1개의 독립변수로부터 종속변수의 값을 예측
		ANOVA 3개 이상 그룹의 평균을 검정	다중선형회귀분석 2개 이상의 독립변수로부터 종속변수의 값을 예측
	질적변수	카이제곱검정 2개 수준의 관계성을 검정	로지스틱회귀분석 독립변수로부터 성공/실패 여부를 예측

  - ex1)
    - 살고있는 지역별로 좋아하는 영화 장르에 차이가 없다/있다 → 지역(질적), 영화장르(질적) → 카이제곱검정
  - ex2)
    - 살고있는 지역별로 영화 장르에 대한 선호도 차이가 있다/없다 → 지역(질적), 선호도(양적) → t검정
  - ex3)
    - 온도, 습도, 위치에 따른 판매량을 예측하려고 한다. → 독립변수 3개, 종속변수 1개, 모두 양적변수에 해당 → 다중선형회귀분석

Section 15

- **카이제곱검정?**
  - 독립변수와 종속변수 모두 질적변수인 경우 사용함
  - 즉, '빈도'과 관찰된 것과 예측된 것이 통계적으로 다른지 검정할 때 사용
- **검정하는 순서**
  1. 각 범주에 대한 기대값(기대빈도)을 구한다.
     - 행 범주 합계와 열 범주 합계를 곱한 후 표본 빈도 총합계로 나눈다.
     - 기대빈도 = 귀무가설이 참이라 가정할 때 모집단의 기대되는 빈도
  2. 각 범주에 대한 관측값과 기대값의 차이를 제곱한 후 기대값으로 나눈다. (카이스퀘어 값)
  3. 이를 합하여 전체의 카이스퀘어 값을 구한다.
  4. 자유도(degree of freedom)을 구한다. df = (행-1)x(열-1)
  5. 유의수준에 해당하는 카이스퀘어 값과 비교하여 검정 결과를 도출
- **카이제곱검정의 2가지 타입**
  - 단, 독립변수와 종속변수 모두 질적변수일 때 연관성을 검증하는 검정임은 동일하다.
  1. **동질성 검정** - **두 개 이상의 모집단**에서 표본추출하여 **각 집단의 범주 비율**을 비교
     - 지역별 모집단에서 호흡기 질환 발병 여부에 대한 검정
  2. **독립성 검정** - **하나의 모집단**에서 표본추출하여 **두 가지 특성(행 범주와 열 범주)** 간 관련성을 비교
     - 특정 모집단에서 자동차 크기와 가족 형태 간 관련성에 대한 검정

Section 16

- **T-검정?**
  - 독립변수가 질적변수이고 종속변수가 양적변수인 경우 사용함
  - 단, 독립변수의 개수가 1개 또는 2개의 그룹의 평균값을 검정
  - ANOVA는 독립변수의 개수가 3개 이상인 경우 사용함
- **검정하는 순서**
  1. 두 집단간 등분산성을 검정한다.
     - 등분산/이분산의 차이에 따른 유의확률 계산 결과값에 대해 차이가 발생

- 카이제곱검정과 마찬가지 흐름이지만, 유의확률 계산 결과 차이가 발생하기 때문에 먼저 검증
- **NOTE: 유의확률 계산에서 독립표본 t검정은 등분산 검정 후 t-검정 실행**
  - 등분산 검정 (F-test) 후 등분산/이분산이냐에 따라 이에 맞는 유의확률을 계산 (계산방법이 달라짐)
  - 두 집단의 평균 차이 검정이기 때문에 집단의 분포의 분산이 다를 수 있고, 그래서 독립표본 t검정에서는 등분산을 확인하는 것이 중요함
  - 아울러, F검정을 사용하는 ANOVA는 t-test와 사실상 동일한 분석방법이나 t검정의 확장임을 참고하자
  - **F-검정 → 두 집단의 분산이 등분산인지 이분산인지 확인하는 검정**
    - 유의수준보다 p값이 크면 귀무가설 채택하는 구조는 동일함 → 즉, 귀무가설 채택 시 등분산
- 단측검정과 양측검정의 직관적 이해
  - 단측검정: 두 집단 간 차이가 있다는 것을 넘어, 둘 중 하나가 더 크다고 주장하는 것에 대한 검정
  - 양측검정: 두 집단 중 어느 한쪽이 크거나 작을수도 있지만 어쨌든 차이가 있다/없다를 주장

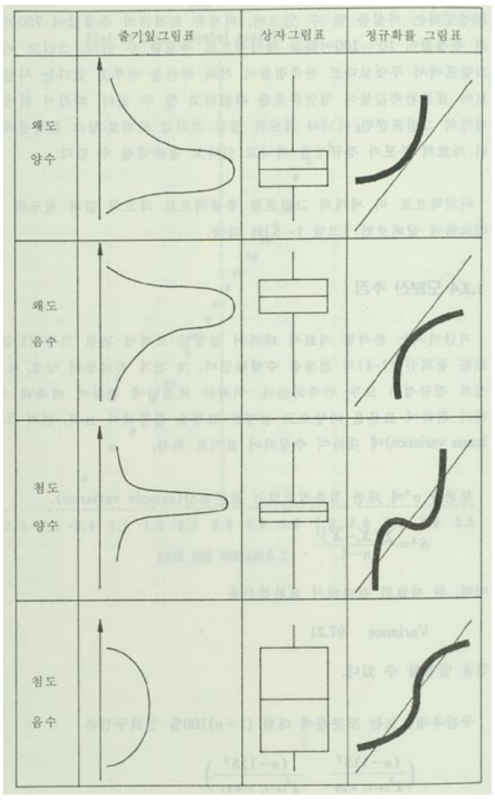
2. 유의수준에 따른 유의확률을 계산하여 결과를 도출한다.

- **1-표본**
  - 하나의 그룹 내에서 단일 변수의 평균과 지정한 상수간의 차이를 검정
    - 즉, 표본평균과 우리가 지정한(가정한) 수치를 비교하며, 지정한 수치일 것이다/아니다로 검정
    - **등분산검정을 할 필요 없이 이분산을 가정하는 이유는, 지정한 상수와 분포간 검정이기 때문!**
      - 모든 표본이 지정한 상수로 통일되어 있는 분포이므로 사실상 수직선
      - **→ 필연적으로 '이분산'일 수 밖에 없음!**
    - 공장에서 생산하는 과자 무게가 100g이 맞는가? → H0 = 100g, H1 ≠ 100g
- **2-표본**
  - **대응표본 (쌍체비교)**
    - 하나의 그룹 내에서 두 변수의 평균 차이를 검정 (같은 대상을 2번 검사함)
    - 1-표본과 다른 점은, 하나의 그룹 내에서 두개의 그룹을 나눈다는 것
      - A마트의 특정 상품의 구매유저에 대해 이벤트 전후 구매량 평균이 차이가 있는가?
  - **독립표본**
    - 두 개의 서로 독립인 그룹의 평균 차이를 검정
    - **등분산/이분산성을 F-검정에 따라 판단한 뒤, 그에 맞는 유의확률 계산법을 따라 수행한다.**
    - 당뇨병 환자를 대상으로 A약을 투약한 그룹과 B약을 투약한 그룹간 혈당 수치 차이가 있는가?

Section 17

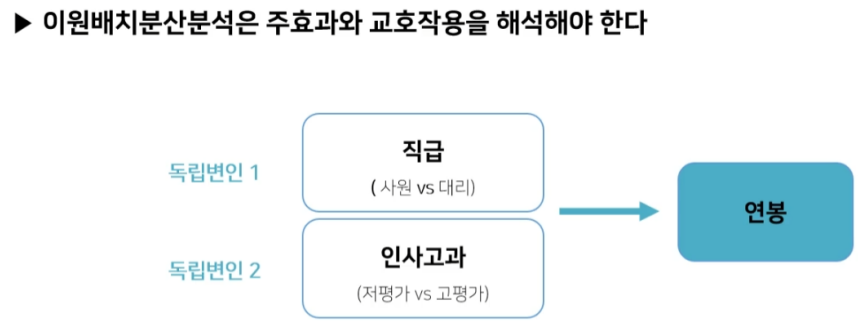
- **분산분석(ANOVA) 이란?**
  - T-검정의 확장된 버전 (평균 차이 검정)
  - 3개 이상의 모집단에 대한 평균 차이를 검정
    - 카이제곱검정은 빈도수에 따른 독립성/연관성을 검증함이 다름 (둘다 유사해보이지만)
  - 전체 표본의 분산과 각 그룹 내부의 분산을 비교하여 검정을 하기 때문에 분산분석이라는 용어를 사용함
  - 그룹을 구분하는 요소 → 독립변인(요인, Factor)
  - 독립변인에 따른 결과를 관찰하고자 하는 요소 → 종속변인
  - 독립변인이 3가지 그룹 이상인 경우 분산분석!
    - 이원분산분석에서 하나의 독립변인이 2가지여도 다른 하나가 3가지 이상이라면 분산분석에 해당
- **분산분석의 종류**
  - **일원분산분석** - 종속변인과 독립변인이 모두 1개
    - 초/중/고 학생간 개인 학습시간의 차이는?
  - **이원분산분석** - 종속변인 1개, 독립변인 2개
    - 남/여, 초/중/고 학생간 개인 학습시간의 차이는?
- **일원분산분석 (일원배치법)**
  - **가설 설정**
    - 요인이 4개면 4개 모두 서로 같다고 귀무가설 가정하고, 적어도 하나의 평균은 다르다는 것을 대립가설로 가정
    - H0: m1 = m2 = m3 = m4
    - H1: 적어도 하나의 평균은 다르다 (즉, H0가 아니다)

- **검정 전 전제 사항 확인**
  1. 변수 설정
     - 독립변수: 지역명 (서울/부산/대전/제주)
     - 종속변수: 사용자 만족도
  2. 정규성 평가
     - 정규성을 띄지 않는 데이터 집합에 대해서는 분산분석을 실행할 수 없음
     - 정규성 평가 방법 중 가장 간단하고 실용적인 방법 → 왜도와 첨도
       - 왜도 → 분포가 얼마나 치우쳐있는가 (비대칭성)
       - 첨도 → 분포가 얼마나 뾰족한가 (이상치)
       - '왜도와 첨도'의 '절댓값'이 특정 범위 안에만 있으면 정규성을 만족한다고 볼 수 있다.
       - 통용되는 일반적 기준
         - **|왜도| < 2~3**
         - **|첨도| < 7~8**
         - 엄격한 정규성 기준은 아니지만, 약간의 왜곡은 허용해주는 최소기준치
       - 왜도와 첨도에 따른 Q-Q플롯과 박스플롯을 통한 정규성 시각화



- **반복없는 - 이원분산분석**
  - 가설 설정이 2가지가 되며, 이 2가지를 각각 검정
    - 지역(독립변인1), 연령(독립변인2) 에 대해 신형 공기청정기에 대한 만족도(종속변인) 차이를 알아보고 싶다.
    - H0: 지역 만족도간 차이가 없다 / H1: H0가 아니다
    - H0: 연령별 만족도간 차이가 없다 / H1: H0가 아니다
  - **검정 전 전제 사항 확인**
    - 일원배치법과 동일하나, 가설이 2개인 점이 차이
    - 변수 설정 시 주의해야함
    - 표로 구성된 데이터에서는 행과 열 2가지가 독립변인이고, 그에 해당하는 값이 종속변인
    - 독립변인1에 대한 분산분석, 독립변인2에 대한 분산분석을 각각 실행하는 것으로 계산됨
    - 즉, p값이 2개가 나오게 되며, 각각마다 귀무가설 채택/기각 하면됨
- **반복있는 - 이원분산분석**
  - 이원분산분석과 동일하지만 데이터에 반복이 생긴다고 하면 사용할 수 있음
    - 데이터의 반복 - 각 요인 조합마다 여러번의 관측치가 존재함
  - 이원분산분석과 달리 독립변인에 대한 상호작용에 대한 가설 설정이 추가되어 총 3개
  - 2개의 독립변인과 종속변인에 대한 평균 차이 검정을 독립변인 개수만큼 검정 (2개)
  - 독립변인 간 상호작용이 있다/없다를 검정 (1개)
    - 이를 **교호작용**이라고 한다.
    - **데이터에 반복이 있는 경우 교호작용이 있는지 추가 확인한다!**

- 예시)
  - 직급, 인사고과에 따른 연봉의 평균차이 검정
  - 가설1: 직급에 따른 연봉차이가 없다/있다
  - 가설2: 인사고과에 따른 연봉차이가 없다/있다
  - 가설3: 직급과 인사고과는 상호작용이 없다/있다
    - 교호작용 측정 가설로, 직급과 인사고과에 따른 연봉차이 검정에서 직급과 인사고과에 따른 연봉 데이터가 여러개가 존재하는 경우 반복있는 이원분산분석이 가능!
- **주 효과와 교호작용**
  - 반복이 있는 이원분산분석에서 알아야할 개념



- **주 효과**
  - 독립변인에 따른 종속변인에 차이가 있는가를 검정한 결과
  - 직급과 인사고과에 따른 평균 차이
    - 직급에 따른 주 효과
    - 인사고과에 따른 주 효과
- **교호작용**
  - 독립변인끼리 서로 영향을 미치는가?
  - 상호작용의 의미와 같음
  - 직급과 인사고과에 따른 평균 차이
    - 인사고과에 따른 연봉차이는 직급에 따라 다르다 / 또는 그 반대

Section 17

- **회귀분석이란?**
  - 연속형 변수 사이의 모형을 구한 뒤 적합도를 측정하는 방법
  - 변수들 사이에서 나타나는 경향성을 설명하는 것이 주 목적
    - 변수들 사이의 함수적인 관련성을 규명하기 위해 모형을 가정하고 이 모형을 측정된 변수의 자료로부터 추정하는 통계적 분석 방법

        - 어떤 관계가 있을지에 대한 여러가지 가설들을 회귀모형이라 부름
        - 수학적으로 표현할 때
            1. 크게 트렌드를 나타내는 부분과
            2. 통제할 수 없는 오차를 나타내는 부분으로 분리됨
    - 활용 가능한 문제들
        - 시간에 따라 변화하는 데이터
        - 어떤 영향을 미치는가?
        - 가설적 실험이나 인과 모델링 등
        - 좌표평면에 표현할 수 있는 데이터들을 이용한 추정에 가장 많이 사용됨
    - 단순선형회귀 (Simple Linear Regression)
        - 두 연속형 변수 간의 관계를 파악하는 것 (x를 이용해서 y를 예측하는)
- **회귀분석 로드맵?**
    1. 단순선형회귀 : y=ax+b
        - x: 키, y: 몸무게
        - 설명변수와 반응변수가 모두 연속형이며, 1개
    2. 다중선형회귀 : y=b0 + b1×1 + b2×2 + ...
        - x1: 키, x2: 허리둘레, y: 몸무게
        - 설명변수가 2개 이상, 반응변수가 1개, 모두 연속형
    3. 로지스틱회귀 : y = (e^(alpha+beta*x))/(1+e^(alpha+beta*x))
        - x: 광고노출시간, y: 클릭여부
        - 설명변수와 반응변수 모두 연속형이지만, 반응변수가 범주형 변수
        - 반응변수 y가 2개 이상인 경우 다중로지스틱회귀를 사용

Section 18

- **단순선형회귀분석**
    - 두 연속형 변수가 관계가 있는지 검정해야함
    - 변수 간 관계가 있는지 검정
        1. **가설 설정**
            - H0: 기온과 아이스크림 판매량은 관계가 없다
            - H1: 기온과 아이스크림 판매량은 관계가 있다
        2. **유의수준 설정**
            - 통상적 기준인 alpha = 0.05로 설정
        3. **회귀분석 실시 (집중적으로 봐야할 것만)**
            - 회귀분석 통계량
                - 다중상관계수, 결정계수, 표준오차, 관측수 등
            - 분산분석
                - 회귀, 잔차에 대한 자유도, 제곱합, F비 등
            - 모수추정표
                - p-값을 살펴봐야 함
                - y절편과 독립변수 2개에 대한 p-값이 계산됨
                - 독립변수가 종속변수와 관계가 있는지 봐야하기 때문에 독립변수에 대한 p값을 확인해야함
- **회귀분석 결과 읽기**
    - 단순선형회귀식 : y=ax+b
        - 단순선형회귀분석을 통해 구해진 기울기 a와 절편 b를 회귀계수라 한다.
    - 회귀계수의 의미
        - y절편(b) : 설명변수가 0일 때, 예상되는 반응변수의 값
        - 기울기(a) : beta계수라고 하며, 설명변수 x가 1단위만큼 증가할 때 예상되는 반응변수의 변화량
    - 가설검정을 통해 두 변수간 상관관계가 있다는 사실은 알아냈음

        - 그러나, 특정 독립변수값에 대해 종속변수 값이 어떻게 되는지도 알아야함
        - 이는 회귀분석식을 통해 알아낼 수 있게 됨
    - 추가적인 정보를 얻는 것
        1. 회귀계수를 가지고 회귀분석식을 완성하기
        2. 특정 독립변수의 값에 따른 종속변수 값 추정하기
        3. 추정값의 신뢰도를 계산하기 (**결정계수/설명력** R^2)
    - 회귀분석표를 계산했을 때, 분산분석과 T검정을 추가로 하는 이유는 모회귀를 추정하기 위해 표본회귀식을 구하고 검정하기 위해서임
        - 결정계수를 구해서 회귀식의 설명력을 알아도, 회귀식이 유의하지 않으면 안되므로 회귀식의 유의성 검정을 위해 F값을 구하는 분산분석을 실시하는 것
        - 회귀계수가 유의한지도 검정할 수 있으며, 이는 회귀식이 의미가 있다/없다로 가설을 설정하여 beta = 0 을 귀무가설로 둔 T검정을 실시한다.
        - 기존에 가지고 있던 표본에서 표본을 새롭게 추가하게 된다면 회귀선도 달라지게 됨
        - 즉 오차가 발생할 수 있고, 이러한 예측값과 실제값에 대한 차이를 **'잔차(Residual)'** 라고 함
            - 또는, 만든 회귀식과 실제값과의 차이라고 표현해도 됨
            - 이 잔차가 커질수록 추정한 결과를 믿기 어려워짐
        - 결정계수는 이러한 설명하지 못하는 부분의 오차(잔차)가 어느정도인지 비율로 표현
            - **결정계수는 확률**이며, 높을수록 추정한 값의 신뢰도가 높다는 의미
            - 평균적인 잔차의 크기를 나타내는 지표 → RMS (Root Mean Square)
        - **잔차의 분포는 정규분포를 따라야한다. 그렇지 않으면 회귀식이 제대로 추정되지 않았다는 의미이다.**
            - +@) 정규분포이기 떄문에 등분산성도 자동으로 만족해야함을 알 수 있다.
            - +@) 표본회귀식의 잔차 epsilon이 최소가 되도록 하면 모회귀식과 점점 같아지게 되므로, 이를 최소로하기 위해 잔차의 제곱합을 최소로 하는 최소제곱법을 이용함
    - 회귀분석 실시 전 반드시 확인해야할 것
        - 쉽게는 결정계수를 확인한다.
        - 깊게는 회귀식과 회귀계수에 대한 유의성 검정을 통해 p값을 먼저 확인해야한다.
    - **추가 학습이 필요한 것**
        - 가정: 선형회귀모형의 오차항은 평균이 0이고 분산이 sigma^2인 정규분포를 따른다.
        - 오차최소화 : 최소제곱법 - 잔차제곱합이 최소가 되도록하는 회귀계수 추정
        - 잔차분석: 추정한 모형이 잔차의 정규성 및 독립성 등을 잘 따르고 있가
- **다중선형회귀?**
    - 단순선형회귀의 확장 버전
    - 가설도 마찬가지로 독립변수마다의 종속변수와의 관계를 가설검정하면 된다.
        - 근무시간, 프로젝트수에 대한 인사고과에 대해 회귀분석은 각 독립변수마다 단순선형회귀에서의 가설검정을 적용하면 된다.
        - 근무시간 vs 인사고과 → 검정
        - 프로젝트수 vs 인사고과 → 검정
- 로지스틱회귀?
    - 연속형 독립변수에 범주형 종속변수에 대한 관계가 있는지 보고 싶은 것
    - 로지스틱회귀를 통해 50%를 기준으로 성공/실패를 나누고 독립변수에 따른 결과를 추정할 수 있음
    - **로지스틱회귀분석 회귀계수의 의미**
        - 오즈비(Odds Ratio) : 성공확률과 실패확률의 비를 의미
            - A와 B의 오즈(Odds)를 각각 구한다.
                - X의 오즈 = X의 성공확률 / X의 실패확률
            - 오즈비
                - 오즈비 = A오즈 / B오즈
            - 오즈비가 1보다 큰 경우 독립변수가 증가함에 따라 사건 발생확률이 증가
            - 오즈비가 1보다 작은 경우 독립변수가 증가함에 따라 사건 발생확률이 감소
        - 오즈에 자연로그를 씌워준 것 = Logit (로짓)
            - 이 로그오즈비는 마찬가지로 0보다 크고 작음에 따라 사건 발생확률이 증가/감소함
        - 따라서 로지스틱 회귀분석은 이 오즈에서 발생하게 된 것
    - **로지스틱회귀식**
        - y = 1 / (1+e^-(y0+beta*x)), y0는 y절편, beta는 beta계수

- 로지스틱회귀분석을 통해 구해진 y0, beta를 회귀계수라 함
- y0 (상수항) : 해석하기 어려움
- beta : 설명변수 x가 1단위만큼 증가할 때 예상되는 반응변수의 로그 오즈비 (log-odds ratio)
- 가설검정에서는 이 beta에 대한 유의확률을 가지고 검정한다.