SQL을 이용한 표본 추출 및 관리 로직 워크플로우 구축하기

1. 목표

조사 응답 데이터를 효율적으로 관리하고, 회차별 표본을 체계적으로 추출·관리 할 수 있는 로직을 만드는 것으로 구체적으로 다음과 같은 목표를 가지고 진행

- N차(1차, 2차, 3차…)에 걸쳐 무중복(random/층화) 표본 추출

-> 이전 회차에 이미 추출된 응답자는 다음 회차 표본에서 자동 제외

- 각 회차별 표본 비율/ 수량의 유연한 변경

-> 회차에 따라서 표본 추출 비율을 조정 가능

- N회 샘플링 후 전체 모집단에서 남은 집합(미배정)군을 즉시 구분

-> 추출 이후에도 전체 모집단과 남은 후보군을 확인 가능

- 모집단(pupulation)은 불변

-> 원본 데이터는 항상 그대로 유지, 추출 결과와 상태는 별도 테이블로 관리

- 재현성 및 속도 확보

-> 동일한 조건과 시드를 주면 언제든 같은 결과 재현, 빠른 검색·추출을 위한 인덱스 활용

2. 표본관리 구축을 위한 엔터티 (데이터 구성 요소)

엔터티(table)명	역할
population	원본 모집단 데이터, 성별, 나이, 지역, 전화번호 등 추출 조건이 되는 속성 포함, 항상 불변으로 유지하기
waves	각 회차별 표본 추출에 대한 메타 정보, 예: 회차명, 추출 방식(랜덤/ 층화), 목표 수, 목표 비율, 사용된 시드값 등
wave_strata_targets	층화 추출 시, 성별·연령대별 목표 수량을 지정하는 보조 테이블
assignments	실제 추출된 표본 결과, 어느 회차에 어떤 응답자가 배정되었는가를 기록, 무중복 보장을 위해 제약 조건을 두기
exclusions	특정 응답자를 영구적으로 제외하기 위한 테이블
views	아직 추출되지 않은 미배정 응답자 집합을 즉시 확인,

3. 운영 흐름

1) 회차 정의

- 새 추출 회차를 waves에 기록 (방식, 목표 수/ 비율, 조건, 시드값 등 메타 데이터 저장)

2) 추출될 후보군 산출

- population 테이블에서 조건(예: 지역=서울, 연령 20~59)을 적용

- exclusions 테이블 및 기존 assignments 테이블 (이미 추출된 응답자) 제외
- 유효 후보군 확보

3) 표본 추출 실행
- 랜덤 추출 : 후보군 전체에 난수를 부여 -> 상위 N명 선택
- 층화 추출 : 성별x연령대 등 층별로 나눈 뒤, 각 층에서 지정 수량만큼 추출
- 추출된 결과를 assignments에 기록 -> 중복 자동 차단

4) 결과 검증 & 로깅
- view 테이블을 통하여 회차별 성별·연령대 분포 검증
- 실행 시점, 조건, 시드값을 waves에 남아 재현성 확보

5) 남은 집합 관리
- 현재까지 추출되지 않은 응답자를 확인
- 차후 회차 표본 추출 대상군으로 활용

4. 무결성 · 성능 · 확인 포인트

- 무중복 보장
- assignments 테이블에서 (wave_id, respondent_id)를 기본키로 지정
- 필요 시, respondent_id 전체에 대해 유니크 제약 걸어 모든 회차 간 중복 차단 가능
- 성능 최적화
- 주요 조건 컬럼(성별, 나이, 지역)에 인덱스 생성
- 대량 데이터에서도 빠른 필터링·랜덤 샘플링 가능
- 재현성 확보
- 회차별 시드값(seed)과 조건(JSON), 실행 쿼리 정보까지 저장 -> 같은 조건이면 언제든 동일 결과 재현 가능
- 원본 불변성
- population 테이블은 원본 그대로 유지
- 변경, 추출, 배정은 모두 파생 테이블 (assignments, waves)에서만 관리

5. 기대 효과
- 투명성 : 표본 추출 과정과 결과를 모두 데이터베이스에 기록
- 유연성 : 추출 회차별로 조건/비율 자유롭게 변경 가능
- 효율성 : 중복 제거, 남은 집합 구분 등 관리 자동화로 시간 단축
- 재현성 : 동일 시드/조건을 주면 언제든 같은 결과 확인 가능

6. ERD 설계도