cM_6 데이터셋: 302개 쌍이 어디로 갔는가?

출처	개수
친족관계 데이터 (원본)	3,786
통계 보유 쌍	3,507
최종 병합 데이터셋	**3,484**
훈련 세트 (80%)	2,787
검증 세트 (20%)	697

계산식: 2,787 + 697 = **3,484** ✓

데이터 격차 설명

격차 1: 친족관계 → 최종 (3,786 → 3,484 = 302개 쌍 손실)

302개 쌍은 친족관계 분류는 있으나 통계 레코드가 없습니다.

이 쌍들은 친족관계 파일에는 존재하지만 분포 통계가 분석되지 않았습니다. IBD 지표는 있지만 백분위수, cM 임계값, 또는 세그먼트 집계를 병합할 수 없습니다.

격차 2: 통계 → 최종 (3,507 → 3,484 = 23개 쌍 손실)

23개 쌍은 통계는 있으나 친족관계 분류가 없습니다.

이 쌍들은 통계적으로 분석되었지만 친족관계 분류가 할당되지 않았습니다. merged_info.out 파일에는 존재하지만 친족관계 파일에 대응하는 레코드가 없습니다.

3,507과 3,484는 왜일까?

병합 과정에서:

- 친족관계 데이터: **3,786개 쌍** (302개는 통계 없음)
- 통계 데이터: **3,507개 쌍** (23개는 친족관계 분류 없음)
- 최종 겹침: **3,484개 쌍** (친족관계 AND 통계 모두 보유)

제외된 쌍 예시 (모두 UN 레이블):

1. `1-1_vs_30-1` — IBD 데이터 있음, 통계 없음
2. `1-1_vs_37-1` — IBD 데이터 있음, 통계 없음
3. `1-1_vs_46-1` — IBD 데이터 있음, 통계 없음
4. `1-2_vs_30-1` — IBD 데이터 있음, 통계 없음
5. `1-2_vs_31-1` — IBD 데이터 있음, 통계 없음 ... (297개 추가 UN 레이블 쌍)

제외된 302개 쌍의 공통 특징:

- **친족관계 레이블:** UN (미분류/모호함)
- **누락된 데이터:** 분포 통계 (백분위수, cM 임계값, 세그먼트 집계)
- **영향:** 모델 훈련을 위해 데이터 품질을 보장하기 위해 이 쌍들을 제외했습니다.