# 산보연 Data check - 6월 9일 Version

작성자 : 이은경

**1. CAL(입사 시기)의 각 범주에 대하여 입사 시기 5년 이후까지의 follow-up year(YEAR 변수)에 대응하는 추적 인년 합계(FY_BZ), 누적 근로자수(ENR_BZ)에 결측값이 있는지 확인.**

: 예를 들어, CAL=0이면 입사 시기가 ~1995년까지이므로, YEAR가 1996년 ~ 2000년 동안 추적 인년 합계, 혹은 누적 근로자 수에 결측값이 있는지 확인

**[Result]**

| CAL 범주 | Follow-up year 범위 | FY_BZ 결측값 개수 | ENR_BZ 결측값 개수 |
|---|---|---|---|
| 0 | 1996년 ~ 2000년 | 116 | 0 |
| 1 | 2001년 ~ 2005년 | 3262 | 0 |
| 2 | 2006년 ~ 2010년 | 4959 | 0 |
| 3 | 2011년 ~ 2015년 | 5476 | 0 |
| 4 | 2016년 ~ 2020년 | 2465 | 0 |

→ "추적 인년 합계" 변수에는 결측값이 있는 것으로 보임 / "누적 근로자 수" 변수에는 결측값이 없음.

**2. YEAR 변수 기준으로 백혈병 발생 건수(LEUKEMIA_BZ) 합한 후, 국립암센터에서 매년 보고되는 백혈병 발병률과 비슷한지 확인.**

2-1) 국립암센터 백혈병 발병률

(참고 link : https://ncc.re.kr/cancerStatsView.ncc?bbsnum=598&searchKey=total&searchValue=&pageNum=1)

| | year | LEUKEMIA_BZ | CR |
|---|---|---|---|
| 1 | 2000 | 200700000 | 4.2 |
| 2 | 2001 | 220900000 | 4.6 |
| 3 | 2002 | 231900000 | 4.8 |
| 4 | 2003 | 228300000 | 4.7 |
| 5 | 2004 | 237200000 | 4.9 |
| 6 | 2005 | 234600000 | 4.8 |
| 7 | 2006 | 244300000 | 5.0 |
| 8 | 2007 | 247700000 | 5.0 |
| 9 | 2008 | 260100000 | 5.3 |
| 10 | 2009 | 271700000 | 5.5 |
| 11 | 2010 | 276000000 | 5.5 |
| 12 | 2011 | 290000000 | 5.8 |
| 13 | 2012 | 287300000 | 5.7 |
| 14 | 2013 | 307200000 | 6.1 |
| 15 | 2014 | 311400000 | 6.1 |
| 16 | 2015 | 328600000 | 6.4 |
| 17 | 2016 | 343700000 | 6.7 |
| 18 | 2017 | 339700000 | 6.6 |
| 19 | 2018 | 352000000 | 6.9 |

LEUKEMIA_BZ : 각 YEAR에 발생한 총 백혈병 발생 건수

**CR(조발생률)** : $\text{조발생률} = \frac{\text{새롭게 발생한 암환자수}}{\text{연앙인구}} \times 100{,}000$

2-2) 산보연 데이터

| YEAR <int> | sum_LEUKEMIA_BZ <int> | sum_FY <int> | rate_per_year_FY <dbl> |
|---|---|---|---|
| 2000 | 121 | 18857990 | 0.642 |
| 2001 | 269 | 22612823 | 1.19 |
| 2002 | 473 | 29265205 | 1.62 |
| 2003 | 680 | 41100139 | 1.65 |
| 2004 | 955 | 51892516 | 1.84 |
| 2005 | 1322 | 65058466 | 2.03 |
| 2006 | 1704 | 78544059 | 2.17 |
| 2007 | 2166 | 93595599 | 2.31 |
| 2008 | 2646 | 109762417 | 2.41 |
| 2009 | 3242 | 126932284 | 2.55 |
| 2010 | 3936 | 145737309 | 2.70 |
| 2011 | 4701 | 165984731 | 2.83 |
| 2012 | 5519 | 187676209 | 2.94 |
| 2013 | 6481 | 210808770 | 3.07 |
| 2014 | 7514 | 234846595 | 3.20 |
| 2015 | 8718 | 261033631 | 3.34 |
| 2016 | 10011 | 289216437 | 3.46 |
| 2017 | 11392 | 319175197 | 3.57 |
| 2018 | 12977 | 350391620 | 3.70 |

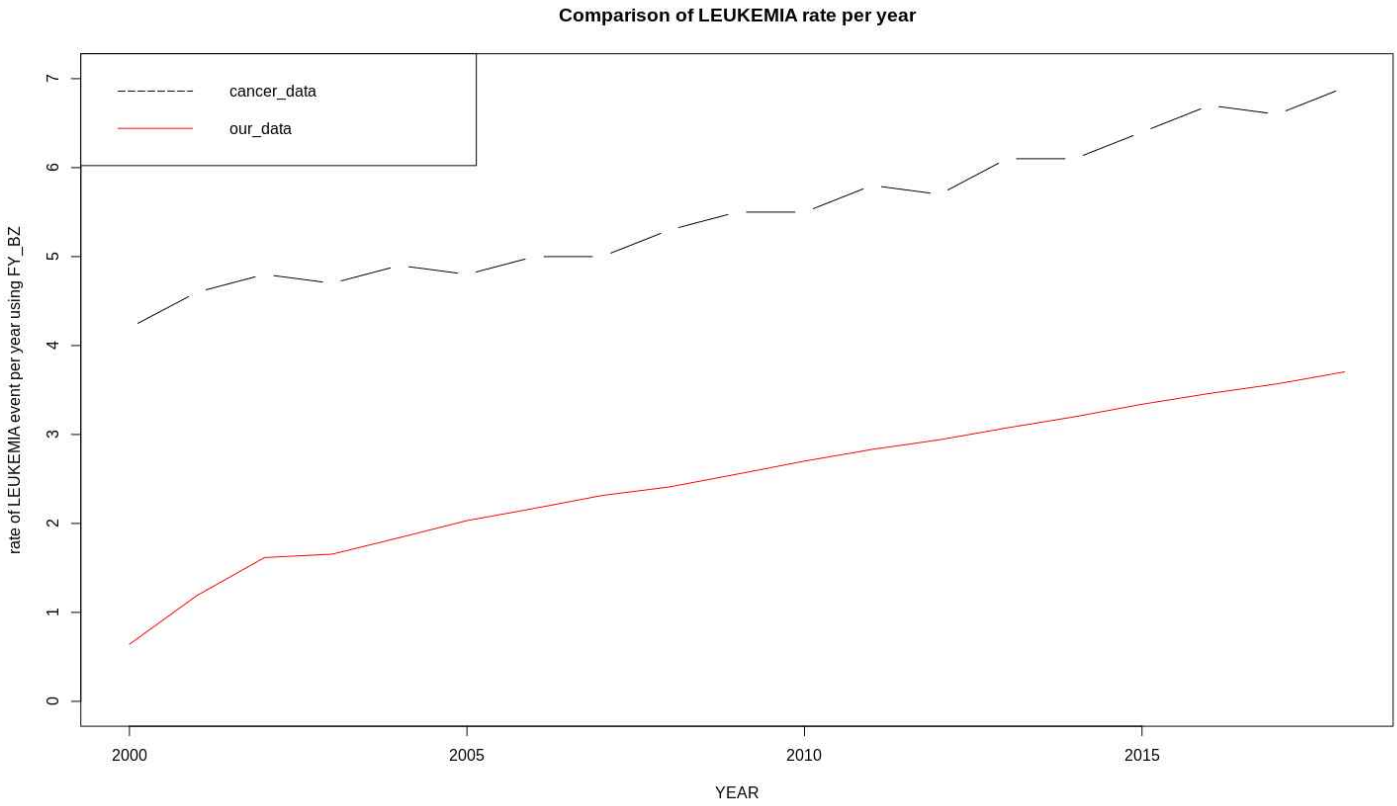
sum_LEUKEMIA_BZ : 각 YEAR에 발생한 총 백혈병 발생 건수

sum_FY_BZ : YEAR에 대응하는 총 추적인년 합계

**rate_per_year_FY** : 위의 CR과 대응되는 개념으로,
(sum_LEUKEMIA_BZ / sum_FY_BZ) * 100000 공식으로 계산됨.

[Graph]

: 빨간색 실선이 our data, **검은색 점선**이 국립암센터에서 보고된 백혈병 추세




→ 국내 총 백혈병 발생 건수가 집계된 국립암센터 데이터값이 더 큰 것이 당연한 부분 / 이때, YEAR에 따른 추세는 비슷한 것으로 보임. (이로 보아, 백혈병 발생률 계산 때, 추적 인년 합계로 나누는 것이 타당해 보임.)

**4. follow up year(YEAR)가 2018년 일 때, 사업장 중분류별 LEUKEMIA_BZ(백혈병 발생 건수), FY_BZ(추적인년합계) TOP 5는 어느 사업장인가.**

1) 백혈병 발생 건수(LEUKEMIA_BZ)

| UP2(중분류) | 업종명 | 발생 건수 |
|---|---|---|
| 41 | **건설업** | 844 |
| 46 | **도매 및 소매업** | 803 |
| 49 | 운수업 | 666 |
| 26 | **제조업** | 648 |
| 75 | **사업시설관리, 사업지원 서비스업** | 551 |

2) 추적 인년 합계(FY_BZ)

| UP2(중분류) | 업종명 | 합계 |
|---|---|---|
| 46 | **도매 및 소매업** | 23213023 |
| 26 | **제조업** | 21837127 |
| 41 | **건설업** | 21328582 |
| 75 | **사업시설관리, 사업지원 서비스업** | 15367009 |
| 86 | 보건업 및 사회복지 서비스업 | 14999701 |

**5. 사업장 중분류 category 기준으로, YEAR에 따른 백혈병 발생률의 변화량 그래프 확인**

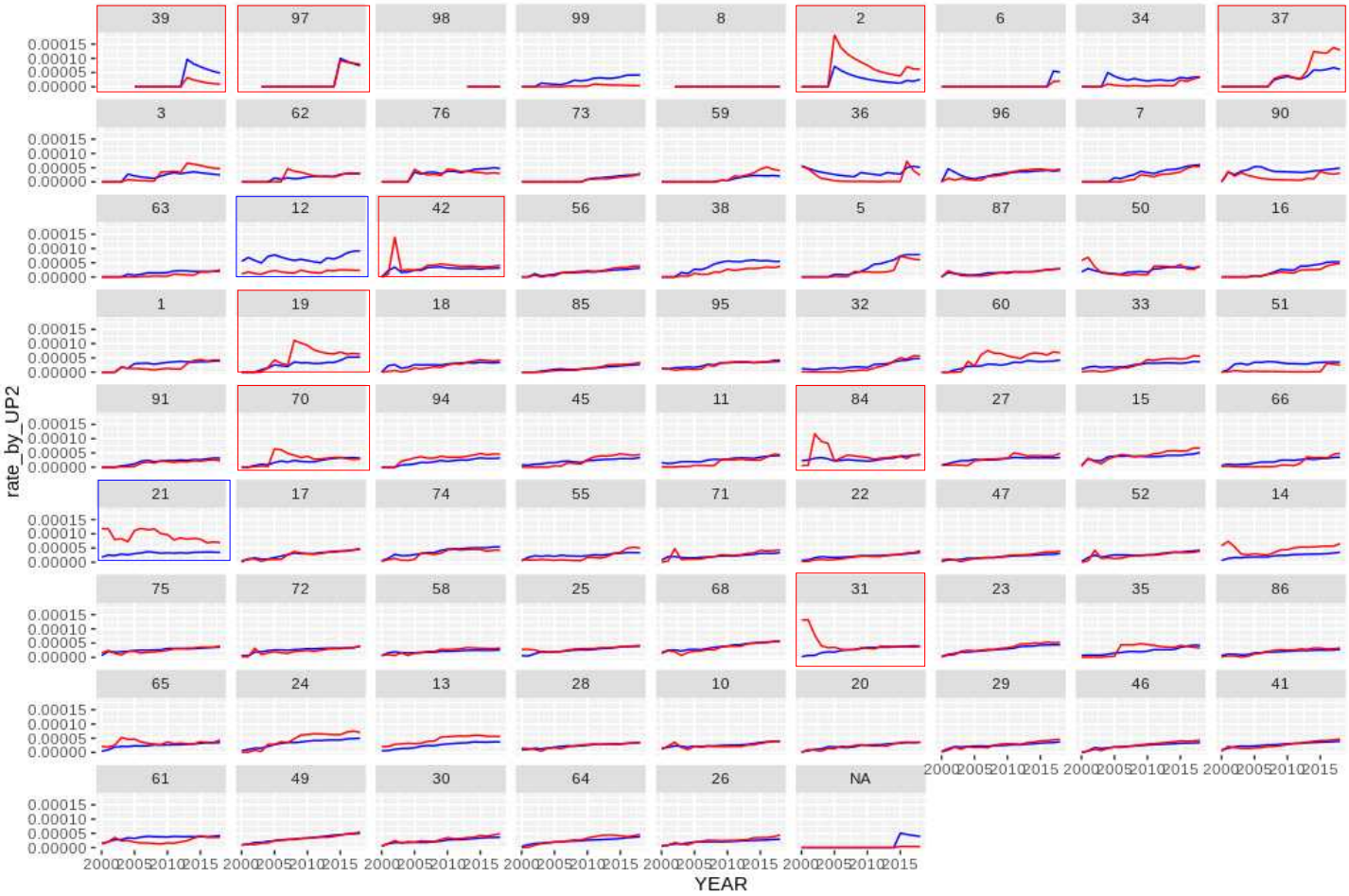
: 백혈병 발생률 계산하는 방법을 2가지로 나눔.

① 중분류별, YEAR로 grouping → 그 후, (백혈병 건수 합) / (추적 인년 합계) 계산

② 각, 사업장별로 rate = (백혈병 발생 건수) / (추적 인년 합계) 계산 → 중분류별, YEAR로 grouping → rate들의 평균 계산

: ①, ② 값을 사업 중분류 별로 겹쳐서 그림. (빨간색이 ①값, 파란색이 ②값 의미)

: 추적 인년 합계 기준으로 plot들을 오름차순 정렬 (plot 순서 = 추적 인년 합계 순서)

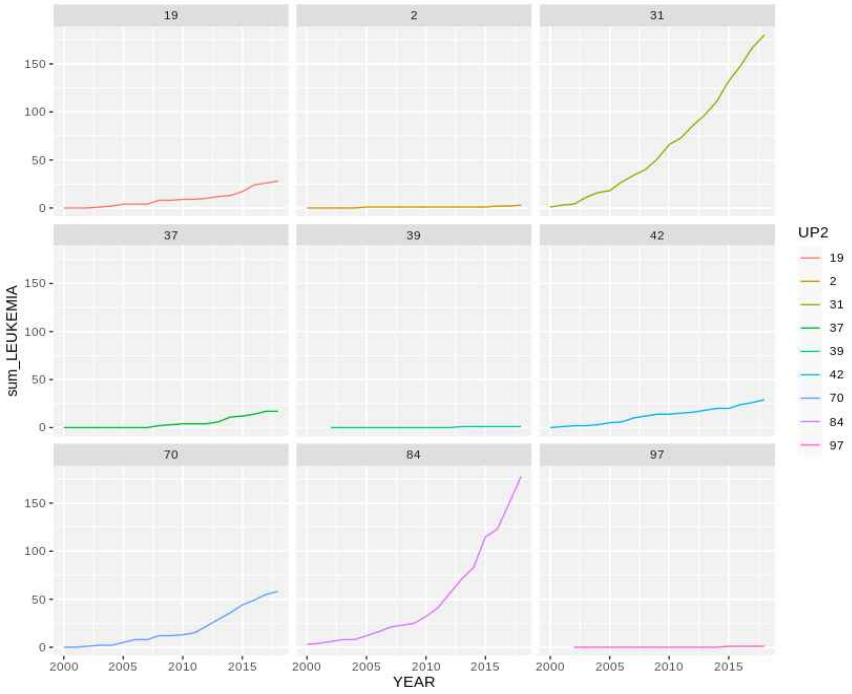



: graph들을 살펴보면, UP2가 "39","97","2","37","42","19","70","84","31"인 사업장은 백혈병 발생 비율이 Peak를 찍는 연도가 있다. ("빨간색" 테두리 참고) / 해당 사업 중분류의 업종명은 아래 표와 같다.

| UP2 | 업종명 |
|---|---|
| 39 | 정화 업(하수, 폐기물) |
| 97 | 가구 내 고용 활동 |
| 2 | 임업 |
| 37 | 하수, 폐기물처리, 원료재생 및 환경복원업 |
| 42 | 건설업 |
| 19 | 제조업(석유, 원유, 코크스 관련) |
| 70 | 전문, 과학 및 기술 서비스업 |
| 84 | 제조업(이동수단 관련) |
| 31 | 공공행정, 국방 및 사회보장 행정 |

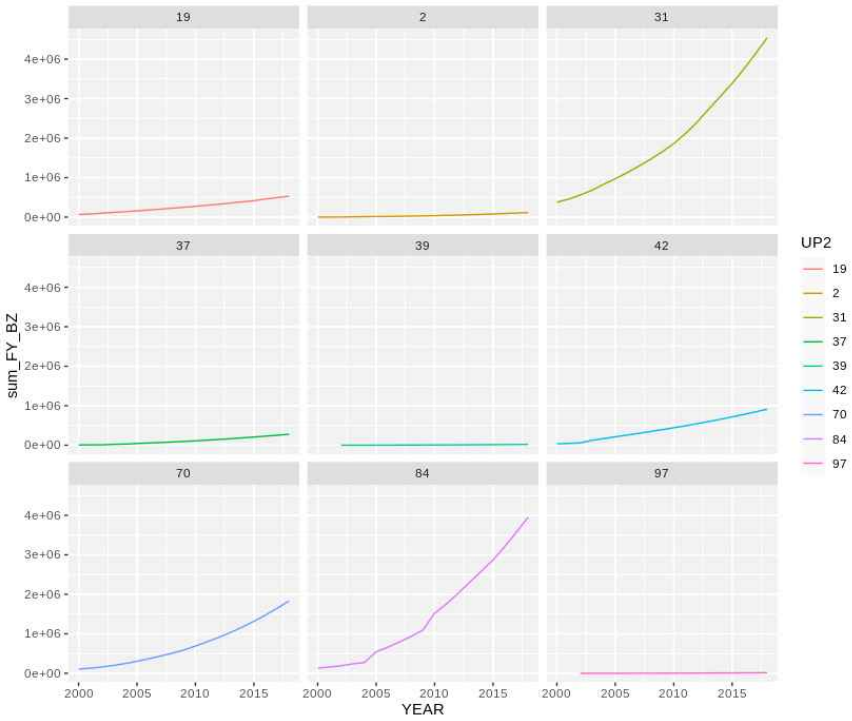
⇒ 이유를 알기 위해, 해당 사업 중분류에 대해서 연도별 백혈병 발생 건수, 사업장 수, 추적인년합계의 변화량을 살펴보았다.

1) 해당 사업 중분류에 대해 연도별 **백혈병 발생 건수**를 시각화해본 결과다.



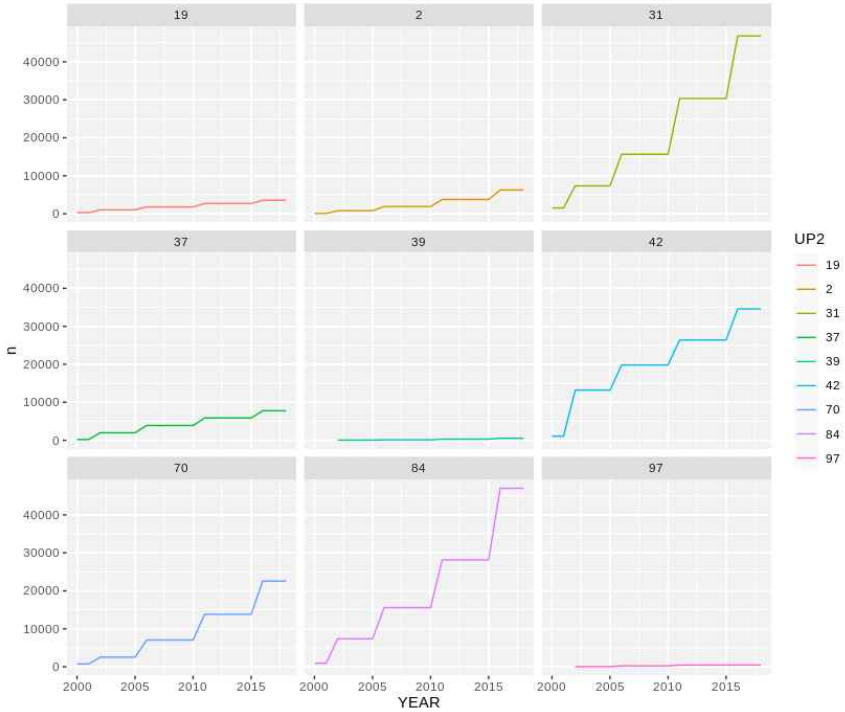

: 사업장이 31, 84인 경우, 백혈병 발생 건수가 가파르게 증가하는 것을 확인할 수 있고, 사업장 42, 70은 가파르지는 않지만, 점진적으로 증가하는 현상을 볼 수 있다.

2) 해당 사업 중분류에 대해 연도별 **추적 인년 합계 변수**를 시각화해본 결과다.




: 1)에서와 마찬가지로, 사업장 31, 84의 추적 인년합계가 가파르게 급증하는 것을 확인할 수 있으며, 사업장 42, 70은 점진적으로 증가하는 것을 볼 수 있다.

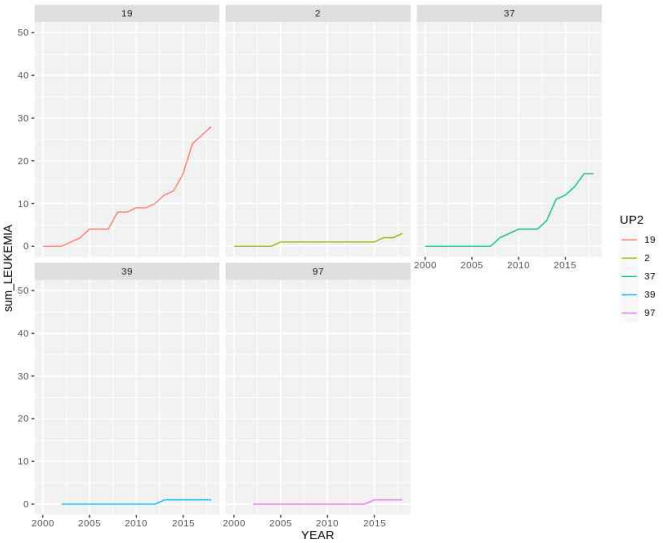
3) 해당 중분류에 대해 연도별 **총 사업장 수**의 변화를 시각화해본 결과다.




: 사업 중분류 31, 42, 70, 84의 총 사업장 수가 급격하게 계단모형을 띄며 증가하는 것을 볼 수 있다.

##########################################################################################

⇒ 이때, 사업 중분류 "39", "2", "97", "37", "19"의 경우는 이유가 뚜렷히 보이지 않아 y축의 범위를 좁혀 다시 살펴보았다.

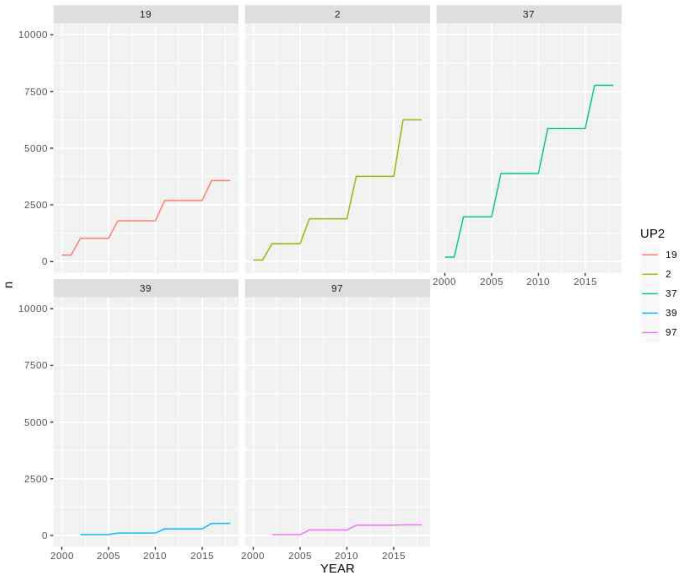


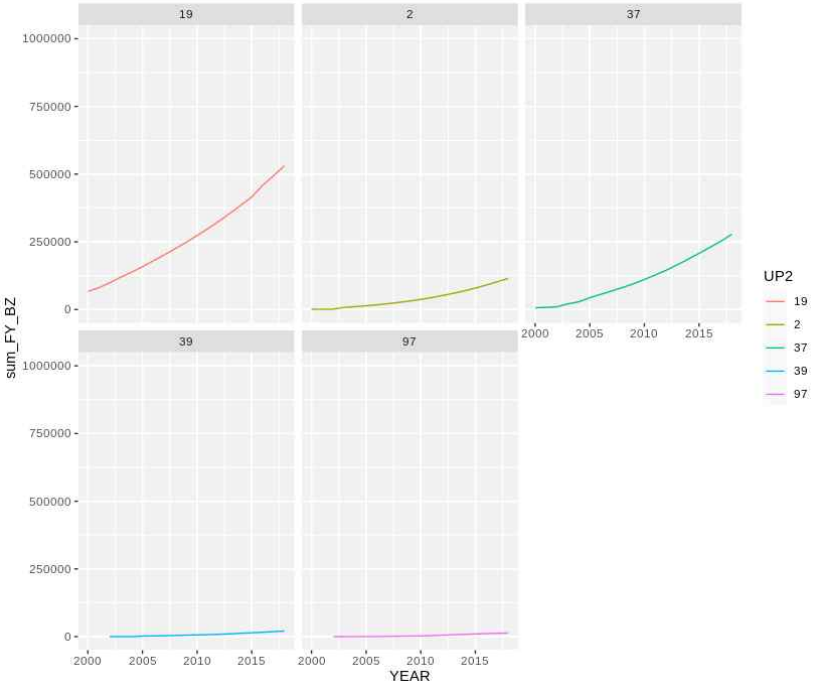

: **백혈병 건수**에 대해 y축의 범위를 좁혀 다시 본 결과, 사업 중분류가 19, 37인 경우 백혈병 건수의 증가가 눈에 띄게 두드러지는 것을 볼 수 있다.

---

---

: **사업장 수**에 대해 y축의 범위를 좁혀 다시 본 결과, 사업 중분류가 19, 2, 37인 경우 사업장 수의 증가가 눈에 띄게 증가하는 것을 알 수 있다.






: 해당 중분류에 대응하는 **추적 인년 합계**에 대해 y축의 범위를 좁혀 다시 본 결과, 사업 중분류가 19, 37인 경우 추적 인년 합계 증가가 눈에 띄게 두드러지는 것을 볼 수 있다.

##########################################################################################

⇒ 결과를 종합해보면 다음과 같다.

| UP2 | 업종명 | 백혈병 발생률이 peak를 찍는 이유 |
|---|---|---|
| 39 | 정화 업(하수, 폐기물) | 추적 인년 합계의 증가에 비해 사업장 수 증가가 더 크다. |
| 97 | 가구 내 고용 활동 | 추적 인년 합계의 증가에 비해 사업장 수 증가가 더 크다. |
| 2 | 임업 | 추적 인년 합계의 증가에 비해 사업장 수 증가가 더 크다. |
| 37 | 하수, 폐기물처리, 원료재생 및 환경복원업 | 백혈병 발생 건수, 사업장 건수의 증가가 추적 인년 합계 증가에 비해 더 크다. |
| 42 | 건설업 | 사업장 수가 급격히 증가 |
| 19 | 제조업(석유, 원유, 코크스 관련) | 백혈병 발생 건수가 급격히 증가 |
| 70 | 전문, 과학 및 기술 서비스업 | 사업장 수가 급격히 증가 |
| 84 | 제조업(이동수단 관련) | 백혈병 발생 건수와 총사업장 수, 추적 인년 합계 동시에 증가하지만, 백혈병 발생 건수와 총사업장 수의 증가가 더 크다. |
| 31 | 공공행정, 국방 및 사회보장 행정 | 백혈병 발생 건수와 총사업장 수, 추적 인년 합계 동시에 증가하지만, 백혈병 발생 건수와 총사업장 수의 증가가 더 크다. |

##########################################################################################

: 한편, UP2가 "12", "21"인 사업장은 백혈병 발생 비율이 다른 업종에 비해 꾸준히 높은 것을 알 수 있다. 해당 사업 중분류의 업종명은 다음과 같다. ("파란색" 테두리 참고)

| UP2 | 업종명 |
|---|---|
| 12 | 제조업(담배) |
| 21 | 제조업(의약품) |