# 2022 2학기 SDA 프로젝트 보고서

2022.12.15. 데이터테크놀로지학과, 김광영(60211642), 송지은(60211670), 신유빈(60211673)

## 초기 문제정의

기획 의도

#### 최근 다시 뉴스나 신문같은 여러 미디어에서COVID-19 우세종에 대한 이야기가 나오고 있다. 팀원 중 김광영의 가족 구성원 중 대다수가 간호사인 관계로 감염병에 대한 관심이 다른 학우들보다 많아 그러한 이슈를 더 관심있게 지켜보게 되었다. 델타 변이가 우세종일 때 보았던 뉴스에서 “바이러스가 진화할 수록 치명률은 낮아지고 감염률은 높아지게 진화한다.”라는 문장이 기억에 남아 데이터 분석을 통해 이 문장이 사실인지 알아보기 위해 실험을 계획하게 되었다.

### 문제 정의

#### 우리 조는 “코로나 바이러스가 진화를 거듭함에 따라 감염률이 증가하고 치명률이 감소한다”는 가정을 증명할 것이다.

#### 현재 한국에서 보인 코로나 바이러스의 우세종은 델타, 오미크론, 그 외 이렇게 세 가지로 분류할 수 있다.

#### 이에 따라 귀무가설(H0)은 “코로나 바이러스가 진화를 거듭해도 감염률과 치명률은 변화가 없다.”로 세우고 이를 기각할 수 있음을 증명할 것이다.

### 단어 정의

#### 우세종은 국내 검출율(국내 + 해외유입사례)이 50%를 넘을 때를 기준으로 파악한다.

#### **감염률(p\_i)는 1일 감염자수/(전체 국민 수 – 전날까지의 누적 감염자 수)로 계산한다. (여기서 i는 infection의 약자)**

#### 치명률(p\_l)은 1일 사망자수/누적확진자수로 계산한다. (여기서 l은 lethality의 약자)

실행 계획

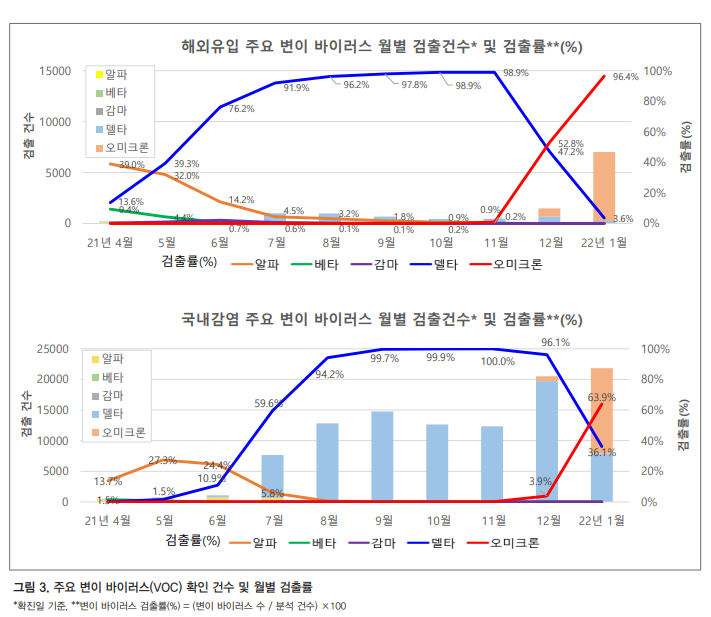
확진자와 사망자에 대한 데이터는 한국의 질병관리청 홈페이지를 이용해서 얻을 것이다.

우세종은 WHO에서 얻은 변이에 대한 정보와 밑의 차트를 참고한다.

#### 22년 1월 24일 – 오미크론 우세종

#### **21년 7월 25~31일 – 델타 변이 우세종**

3개의 집단으로 분류됨으로 ANOVA를 사용해 집단 간 평균의 다름을 증명하고, 이후 2개 집단에 대한 비교를 진행 해 감염률의 상승과 치명률의 감소를 볼 것이다.



팀프로젝트 결과 보고서

단어 정의 수정

#### 설계를 하고 분석을 실행하면서 치명률의 정의에서 2가지 문제가 발생했다. 첫 번째 문제는 처음 정의한 치명률의 정의에서 야기됐는데, 코로나 발병 초기에서 치명률이 너무나 작아 검정의 진행되지 않았다. 따라서 이는 한겨레의 미래&과학 기사(https://www.hani.co.kr/arti/science/science\_general/964555.html)에 나온 기사의 치명률 정의를 따르기로 했다. 두 번째 문제는 사망자의 사망원인이 사망자로 집계된 당일의 우세종에 영향을 받았나 하는 사실이다. 그래서 우리는 이를 17일 이전에 확진자에서 사망자가 나올 확률이 높다고 이야기했다.

#### 치명률 : (누적 사망자 수 / 17일 전 누적 확진자 수)로 계산한다.

진행 과정

#### 데이터를 먼저 날짜를 기준으로 3가지 군으로 분류했다. 여기서는 오미크론, 델타, 그 외 변이로 분류했다.

#### 이후 감염률, 치명률을 위에서 내린 정의에 따라 구하고 신천지 집단 감염 사태의 데이터를 outlier처리해 데이터에서 제외했다.

#### 이후 정규성 검증인 Shapiro test와 등분산검증인 fligner test를 진행했다.

#### 세 집단의 평균의 다름을 증명해야 하기 때문에 Kruskal-Wallis Test를 진행해 다름을 밝혔다. 위에서부터 그 외, 델타 변이, 오미크론 순이다.

#### 텍스트이(가) 표시된 사진 자동 생성된 설명텍스트이(가) 표시된 사진 자동 생성된 설명

#### 위의 결과를 살펴보면, delta, omi, etc 모두 정규성을 띄지 않고, 등분산성을 가지지 않는다. 또한 Kruskal Wallis test의 p-value가 적정 유의수준 0.05보다 작아서 귀무가설인 각 군의 평균이 같다를 기각한다. 따라서 감염률과 치명률의 각 군의 평균이 적어도 하나 다름을 관찰 할 수 있다.

#### 텍스트이(가) 표시된 사진 자동 생성된 설명마지막으로 각각의 군끼리 rank-sum test를 실행한다. 각각 군들의 p-value가 모두 유의수준 0.05보다 작으므로 귀무가설을 기각한다. 따라서 각 군들의 평균은 다르다는 결론을 얻을 수 있다. bartlett test를 통해 유의수준을 0.05/3까지 낮췄어도 귀무가설을 기각 할 수 있다.

설계 평가 및 문제점

실험의 데이터, 실험 설계 문제가 있었다고 생각한다.

#### 이 실험에서 유의미한 데이터를 얻어 내고 싶었으면, 코로나 상황 초기에 있었던, 구로 콜센터, 신천지, 이태원 클럽 이 세가지 집단 감염에서의 n차 감염자들의 데이터를 모두 제거해야 했다. 즉, outlier의 정의와 처리가 매우 미숙했다. 처음에 예상 했던 "치명률은 변이가 진행할 수록(코로나 발병 후 시간이 진행 될 수록) 감소한다"와는 다른 구간이 4곳이나 보인다.

#### 또한 우리가 SDA 수업시간에 배운 다중 선형 회귀 분석까지의 내용은 처음에 설계한 실험 목적인 "k 배의 감염률을 가지고 p 배의 치명률을 가진다"를 유도하기에는 충분하지 않았다.

#### 또한 분석 과정에서 우리가 배운 정규성검정이나 등분산성검정은 유의미하지 않았던 실험이라고 생각한다. 실험에서 증명을 원하는 데이터는 각 군별로 감염자 비율은 증가해야 했고, 치명률은 계속 낮아져야 했다. 따라서 각 군들은 정규분포를 따를 수 없다.

#### 특정 변이가 우세종이라고 해서, 모두 그 변이의 환자라는 가정을 실험을 진행할 때는 인지하지 못하고 기저사실로 실험을 진행했는데, 정확한 데이터를 살펴보기를 원했으면 각 변이의 비율을 실험을 진행할 때 반영해야

#### 했

#### 다

#### .

추가 분석 : 집단 감염자 제외 시도

#### 전처리 : 신천치, 콜센터, 클럽 관련 확진자 제외

#### 데이터가 정제된 pc\_res의 일일 감염자 수를 구해 원래 우리가 사용했던 데이터의 알맞은 날짜까지 교체했다.

#### 데이터가 사망자 수는 구할 수 없다, 하지만 감염자수가 변경되면 치명률도 이 데이터에 맞게 수정될 것이 다라고 예측하고 실험을 진행한다.

#### 치명률의 plot이다. 제일 처음 그래프가 신천치 데이터를 제거 하고 실행했던 처음 실험이다. 두 번째 그래프는 이상치 제거 후 그래프이며, 마지막 그래프는 이상치 제거가 없는 그래프이다.

#### 

#### 

#### 

#### 서로 같은 분포인지 알아보자

#### .텍스트이(가) 표시된 사진 자동 생성된 설명

#### 치명률은 이상치를 제거하지 않은 데이터와 비교했을 때, p-value가 유의수준 0.05보다 작기 때문에 다른 분포를 가지고 있음을 알 수 있다.

#### 감염률도 살펴보았다. (위에서부터 신천지 제거, 이상치 제거, 이상치 제거 없는 그래프)

#### .

#### .

#### .

#### 위의 plot을 살펴보면 별로 큰 차이를 느끼지 못했지만, 명확히 수치화 된 데이터가 없기 때문에 실험을 계속 진행하였다.

#### 텍스트이(가) 표시된 사진 자동 생성된 설명

#### 감염률은 두 비교 모두 p-value가 유의수준 0.05보다 충분히 크다. 따라서 모두 같은 분포를 따른다.

#### .

#### 이상치를 제거하지 않은 데이터와 이상치를 모두 제거한 데이터를 비교했을 때, p-value가 유의수준 0.05보다 작으므로 서로 같은 분포가 아니다.

#### 따라서 이상치를 제거한 실험을 다시 진행한다. 감염률, 치명률 plot이다. 순서대로 etc, delta, omi 변이 순이다.

#### .

#### .

#### .

#### 

#### 

#### 

#### .

#### 

#### 실험의 결론은 위의 첫 번째 실험의 결론과 같다.

데이터 참조

#### 데이콘의 포스트 코로나 데이터 시각화 경진대회에서 집단 감염자 제외한 코드를 copy해 진행했다.

#### https://dacon.io/competitions/official/235618/codeshare/1501?page=1&dtype=random

#### 추가 데이터는 https://www.kaggle.com/datasets/kimjihoo/coronavirusdataset 에서 가져왔다.

회고

#### 교수님이 쓰레기 데이터로 분석을 시작하면 쓰레기가 나올 수 없다 라는 말을 몸소 체험한 분석이었다.

#### 치명률 정의에 너무 매몰되어서 실험은 진행한 것이 실패를 하는 중요한 요인이었다.

#### 치명률 정의에 너무 시간을 낭비해 결국은 시간에 쫓겨 마감하고 이상한 결론이 나올 수 밖에 없었다.

깃허브 링크

https://github.com/Jieun-Song/2022-2-SDATeamProject

감사합니다.