코로나 전/후의 대여소별 자전거 대여량에 대한 베이지안 다중 검정

이동균

December 14, 2021

목차

1. 문제 설명

- 2020년 2월 이후의 집단적인 코로나 감염과 3월 초의 위기경보 격상을 통해서 코로나에 대한 위기감을 인지한 사람들의 외출 자제로 인한 자전거 대여량의 차이를 대여소별로 베이지안 다중검정을 실시하여 차이를 파악해본다.

- 코로나 바이러스 발생 이후 대여소별로 자전거 대여량이 얼마나 차이가 있는지를 알아볼 수 있다. (여기서는 2019년 $3 \sim 5$월의 대여량과 2020년 $3 \sim 5$월의 대여량을 비교하였다.)

1. 문제 설명

- 검정하고자 하는 문제 : 대여소 별로 일별 대여량의 차이의 평균에 대한 사후 분포를 구하고, 검정

- 사전분포 : 1421개 대여소 별 대여량 차이의 평균을 구하고, 그 평균들의 분포를 사전분포로 정의
- 관측치 : 대여소 별로 일별 대여량의 차이(2020년 해당 날짜의 대여량 - 2019년 해당 날짜의 대여량) 관측치가 된다.

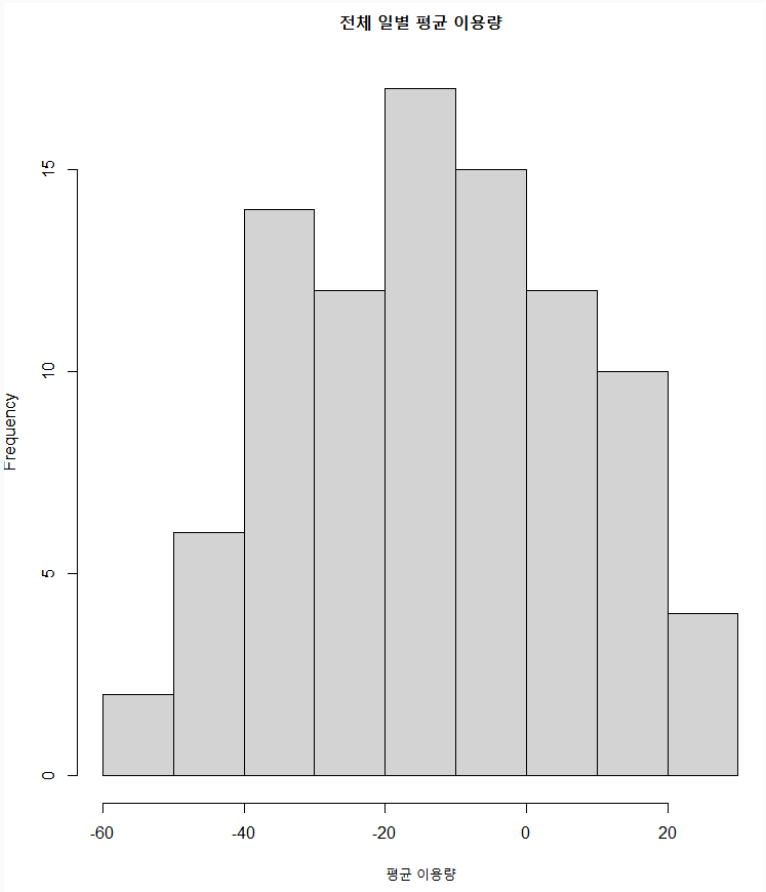
2. 데이터 전처리

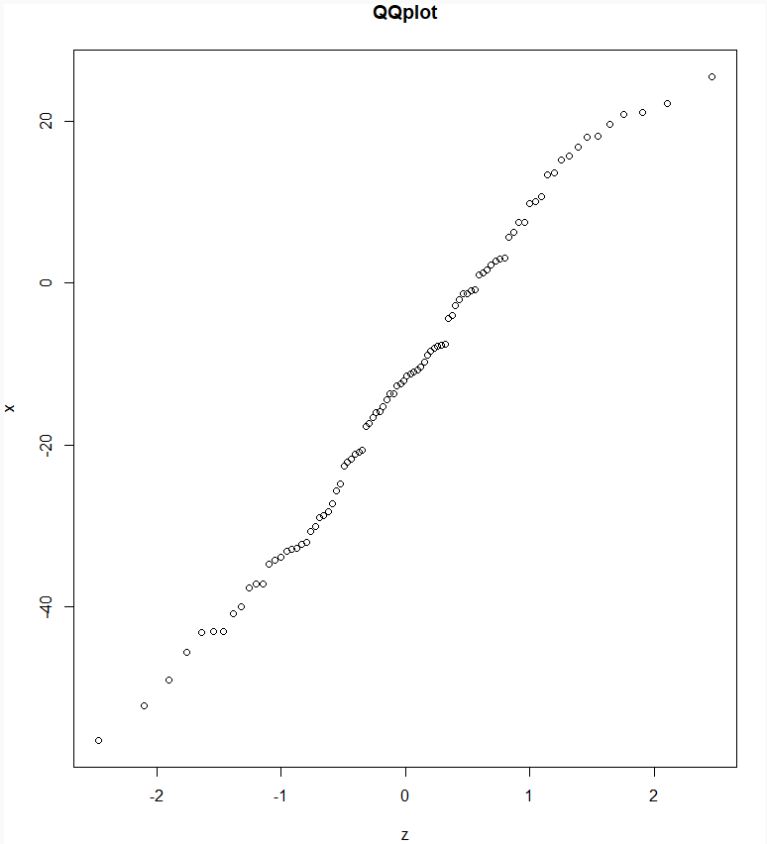
1. 서울시 열린 데이터 사이트의 서울시 공공 자전거 서울시 공공자전거 이용현황에서 대여이력정보 데이터를 수집
2. 대여소별로 일일 대여량의 차이(2020년 날짜의 대여량- 2019년 날짜의 대여량)를 Column 값으로 가지도록 R 프로그램을 이용하여 전처리

Index	관측소 번호	위도	경도	03/01	03/02	···	05/31
1	301	37.57579	126.9715	-1	-27	···	-33
2	302	37.57595	126.9741	-20	-34	···	-102
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1421	1062	37.55108	127.1626	6	0	···	-9

3. EDA

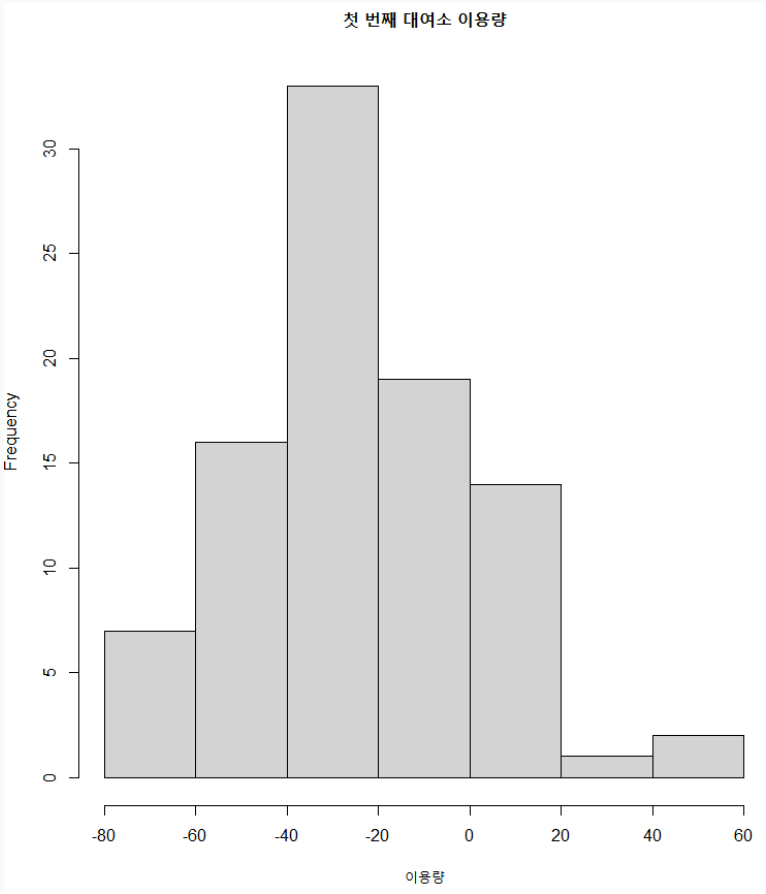
- 서울시 일별 평균 대여량 차이의 히스토그램 및 QQplot

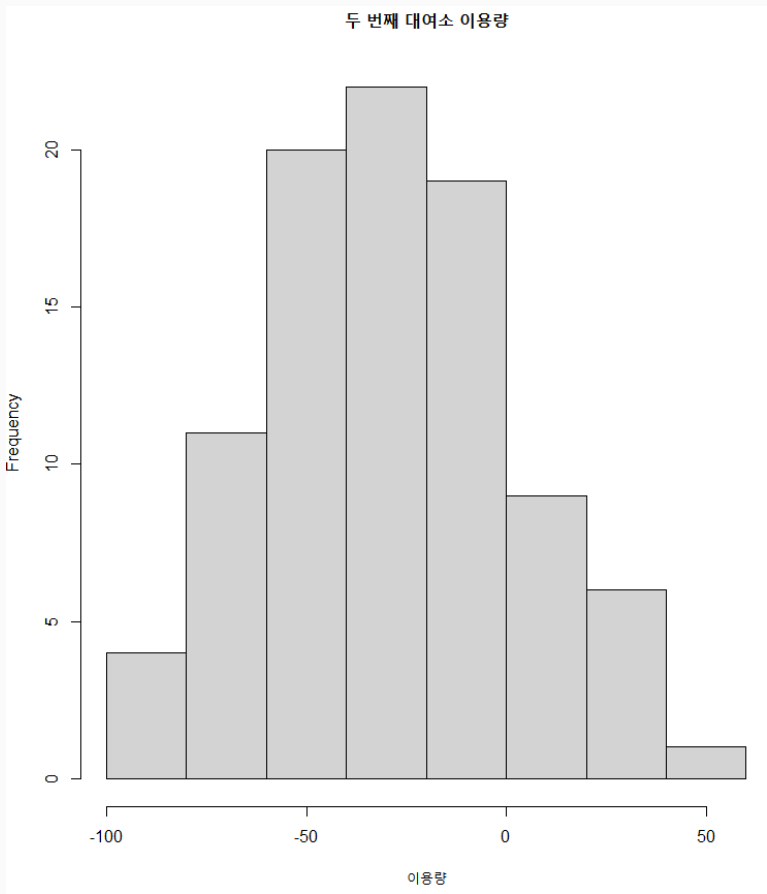




3. EDA

- 대여소별 일별 대여량 차이의 히스토그램





4. 분석 방법 및 결과

각 대여소별로 관측치와 사전분포는 정규분포이고 관측치의 분산(σ^2)을 안다는 가정 하에 사후분포를 구해서 검정을 실시한다.

$$\overline{X}|\theta \sim N(\theta, \sigma^2/n)$$
$$\theta \sim N(\mu_0, \sigma_0^2)$$
$$\theta|\bar{x} \sim N\left(\frac{\frac{1}{\sigma^2/n}\bar{x} + \frac{1}{\sigma_0^2}\mu_0}{\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}}, \left(\frac{1}{\sigma^2/n} + \frac{1}{\sigma_0^2}\right)^{-1}\right)$$

- θ 사전분포의 평균과 분산은 서울시의 대여소별 일별대여량 차이의 평균값들의 평균과 분산으로 가정하였다.
- 대여소별로 분산을 추정하여 σ^2 에 해당하는 값으로 가정하였다.

4. 분석 방법 및 결과

- Parameter 값 구하기

사전분포의 평균(μ_0) : -12.63978
분산(σ_0^2) : 154.2494
n : 92

Index	관측소 번호	위도	경도	관측치 평균	관측치 분산	사후분포 평균	사후분포 분산
1	301	37.57579	126.9715	-19.4891304	449.17570	-18.8605004	4.7325487
2	302	37.57595	126.9741	-34.3260870	1172.04634	-22.4172712	6.5546430
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1421	1062	37.55108	127.1626	-0.3478261	27.96560	-0.3451752	0.3033761

4. 분석방법 및 결과

각 대여소별로 가설은 아래와 같이 설정하고, 각각의 사후확률을 구하여 가장 높게 추정되는 값에 해당하는 가설을 채택한다.

$$H_1 : \theta < -12.63978$$

$$H_2 : -12.63978 \leq \theta < 0$$

$$H_3 : \theta \geq 0$$

4. 분석방법 및 결과

- 각각 대여소별로 사후확률이 높은 가설을 하나씩 채택한다.

Index	관측소 번호	위도	경도	p1	p2	p3	채택
1	301	37.57579	126.9715	0.98863	1.137037×10^{-3}	0	1
2	302	37.57595	126.9741	1	0	0	1
3	303	37.57177	126.9747	0.9999739	2.611479×10^{-5}	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1420	1061	37.54693	127.1346	3.742430×10^{-6}	0.9999402	5.610641×10^{-5}	2
1421	1062	37.55108	127.1626	3.38965×10^{-110}	0.752859	0.2497141	2

4. 분석방법 및 결과

대부분의 대여소가 첫번째 혹은 두번째 가설을 채택하므로 대부분의 대여소에서 대여량이 줄어들었다고 볼 수 있다. 그리고 영등포구, 마포구, 중구, 광진구 등의 지역에서는 서울시의 평균보다도 더 많이 대여량이 줄었다고 볼 수 있다.

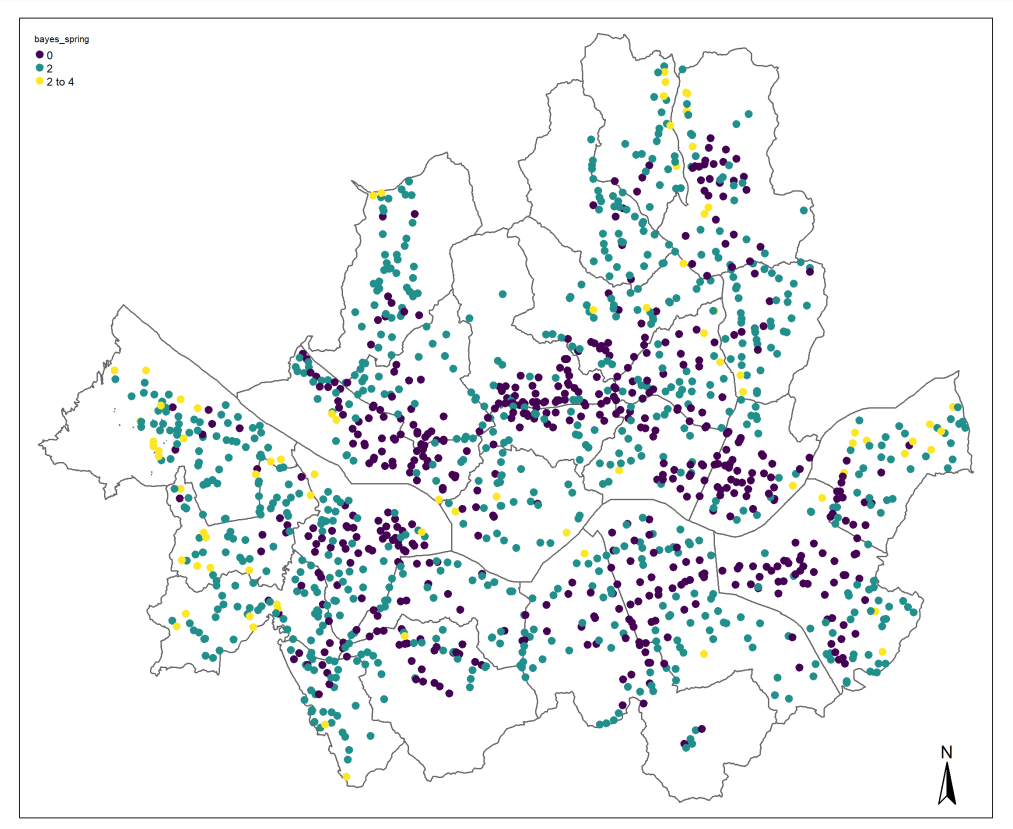


Figure 1: 보라 : 1번째 가설 / 청록: 2번째 가설 / 노랑 : 3번째 가설