A5: Extension Plan
DATA 512
Laura Thriftwood
November 10, 2021

Motivation/problem statement:

The COVID-19 pandemic disrupted existing human behavior at a magnitude previously unseen in modern history. While much of the impact to daily life was negative, there were a number of positive impacts that resulted from the large-scale change to daily human behavior. One such change was the massive shift to remote work at the beginning of the pandemic. While many have since returned to the office, a large number of people will remain partially or fully remote permanently. A side effect of this behavioral change was a decrease in the number of people commuting to work and an overall reduction in traffic. Drastically fewer cars on the road would ideally result in fewer traffic accidents. This assumption is what I would like to explore, with a focus on Salt Lake County, Utah.

I would also like to explore whether the type and severity of crashes differ significantly before and during the pandemic. While fewer cars on the road might make collisions less likely to occur, it's possible that the cars in lighter traffic travel at a higher speed which could cause more severe injury in any resulting crashes. This change in human behavior (driving faster on an open road) is worth exploring with the available data.

If possible, I'd also like to investigate whether the progression of infection rates during the pandemic as explored in A4 has any correlation with a change in the location of crashes, perhaps related to increased travel to hospitals in the area. This is a stretch goal as I'm unsure if the available data will support his type of analysis.

Research questions and/or hypotheses

Did the reduction in traffic levels due to a large-scale shift toward remote work during the COVID-19 pandemic result in significantly fewer vehicle crashes in Salt Lake County, Utah? Did the type or severity of vehicle crashes during the pandemic differ significantly from those that occurred prior to the pandemic?

Data used

The dataset I will use comes from the Utah Department of Public Safety which provides a Raw Crash Data[1] in a .csv file with data as recently as a few days ago. As stated on their website, "[t]he data for the Utah Crash Summary is derived from Utah crash reports. These reports are completed by law enforcement officers throughout the state who investigate crash scenes on

[1] https://udps.numetric.net/utah-crash-summary#/?view_id=7

public roadways. Information is collected when a crash involves injuries, deaths, or at least $2,500 property damage. Crash reports are uploaded daily to the Utah Department of Public Safety data warehouse for central collection. Additional information is collected on fatal crashes at the Department of Public Safety's Highway Safety Office and compiled into the Fatality Analysis Reporting System (FARS) database. FARS is a national data system containing data on all fatal traffic crashes in the U.S."[2]

I do not see any ethical concerns regarding the use of this dataset. The dataset is public and does not contain any identifiable information and the collection method appears sound and unbiased.

This dataset will allow me to analyze crash frequency and details related to each crash event, with data from 2010 to present. The datetime stamp will support the level of analysis required to perform the hypothesis test described in the Methodology section below. It also provides supporting information like the latitude and longitude of each incident, categorical descriptors related to various features of each crash event, such as the road condition and crash severity. The lat/long coordinates will support alignment of this data with other datasets for additional analysis.

Field Name	Description	Data Type
Crash Id	Unique identifier	string
Crash Date & Time	Date of the incident	datetime
Year	Year	int64
Milepoint	Milepoint of incident	float64
KABCO Crash Severity	Categorical description of crash severity	string
Manner of Collision	Categorical description of the collision	string
Roadway Junction/Feature	Categorical description of the roadway	string
Lighting Condition	Categorical description of lighting	string
Weather Condition	Categorical description of weather	string
Roadway Surface Condition	Categorical description of roadway surface	string
Number Vehicles Involved	Count of vehicles involved in incident	float64
Roadway Description-Array	Array of categorical descriptions of roadway	array

[2] https://highwaysafety.utah.gov/crash-data/

County Name	County name	string
City	City name	string
Lat	Latitude coordinate	float64
Long	Longitude coordinate	float64

To explore the crash location data as it relates to hospital location, I may need to supplement my research with an available dataset of "Utah Hospital Characteristics"[3] from the Utah Office of Health Care Statistics or a similar dataset from the Wasatch Front Regional Council[4]. The second dataset contains information on Utah hospitals including both location (lat/long) and other feature data like the number of beds and trauma level. Both datasets include county data which will make it straightforward to align with the crash data.

Unknowns and dependencies

As this is an observational study rather than a randomized experiment, my interpretation of the results of the hypothesis test will be different due to the potential bias inherent to observational studies. I will need to explore possible statistical analysis methods to address this bias, such as stratification.

I'm not confident that an analysis of the crash location data or hospital location data will reveal any usable or useful information at this point but has the opportunity to provide additional insight or context.

Methodology

The crash data will need to be filtered to only include crashes in Salt Lake County and may have other minor cleaning required to prepare it for analysis. As the dataset contains the date/time stamp for each incident, I do not foresee any issues with a comparison with the time series analysis in A4.

I will use hypothesis testing to compare the two population means. As the dataset provides large enough sample size for both populations, I will likely use a large-sample Z-test. This also does not require an assumption of equal variances. The populations are independent from each

[3] https://opendata.utah.gov/Health/Utah-Hospital-Characteristics/ierb-h3t5

[4] https://data.wfrc.org/datasets/utah-hospitals/explore?location=8.367254%2C-15.457895%2C2.09&showTable=true

other as the crash behavior of the pre-pandemic population does not have any impact on the crash behavior during the pandemic.

I will define the null hypothesis as

$$H_0 : \mu_A = \mu_B$$

where μ_A and μ_B are the mean number of crash incidents in the county for time period A (5 years prior to the pandemic) and period B (during the pandemic). μ_A is considered the mean number of crashes of the population.

I will set the alternative hypothesis to be

$$H_1 : \mu_A \neq \mu_B$$

I plan to supplement this analysis with a visualization that overlays the infection rate data from A4 with the crash data on a shared x axis timeline. The dual y axes would show the changing rate of infection and the number of crashes, perhaps smoothed with a rolling average similar to the infection rate time series.

For the secondary questions about the type/severity of crashes, simple visualization methods may be more appropriate. For example, a simple comparison of a stacked to 100% bars comparing the proportions of each categorical variable associated with crashes before and during the pandemic might suffice. If there is time, I'd like to attempt a pairwise comparison of the categorical variable descriptors, perhaps with ANOVA test to compare the variance between the two populations.

Timeline to completion

- Acquire and load the data sets - Complete
- Light EDA and proof of concept - Complete
- Clean and prepare the data – 11/14/21
- Additional EDA - 11/17/21
- Calculate metrics for hypothesis testing – 11/21/21
- Run hypothesis test – 11/28/21
- Evaluate the results and rerun with any modifications – 12/1/21
- Create visualization of the results – 12/5/21
- Prepare presentation (A6) – 12/7/21
- Write final report (A7) – 12/14/21