External Validity Write-Up (Plan 3)

EMSE 6577

Drew Gobbi, Olivia Wang, Olatunji Akinbule, Maddie Warndorf

**1. Identify a source of pilot data that you will use to answer your Question.**

The source of our pilot data is from the feature layer information published by TempeData on Esri ArcGIS. This data supports the Opioid Wastewater Collection Data Dashboard (<https://arcg.is/ey0Ha>). The link for finding the data is below along with the GitHub repository that shows how the data was pulled.

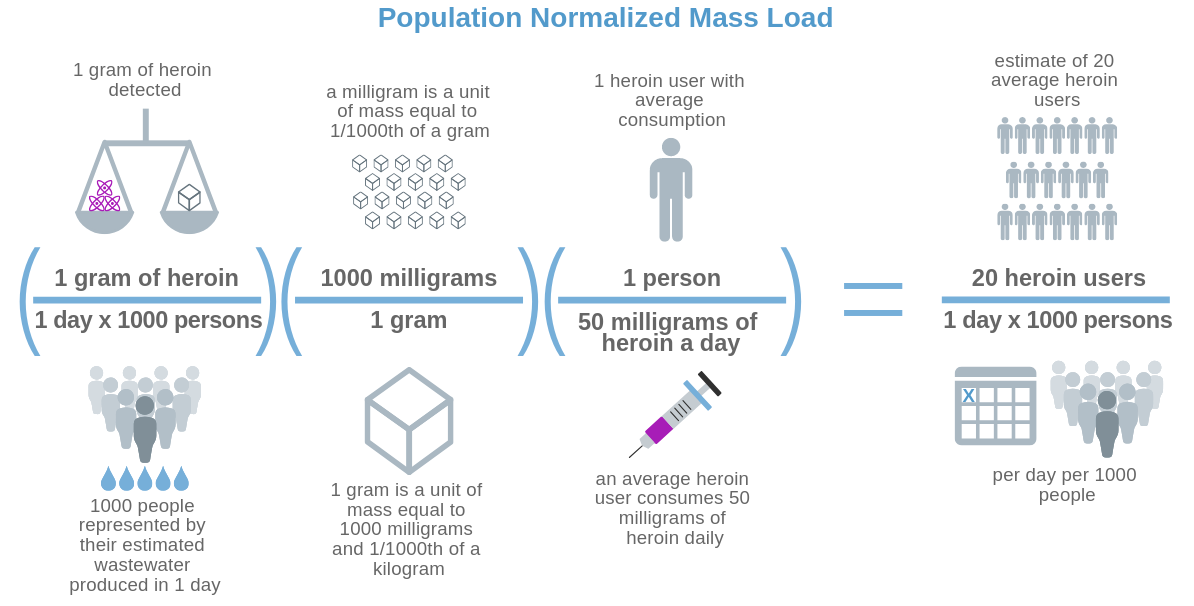
Data Website: <https://www.arcgis.com/home/item.html?id=93ee24cbf945405c8469742555670fc5>

GitHub Repository:

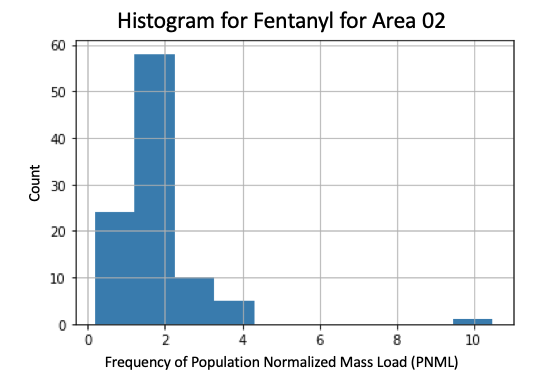
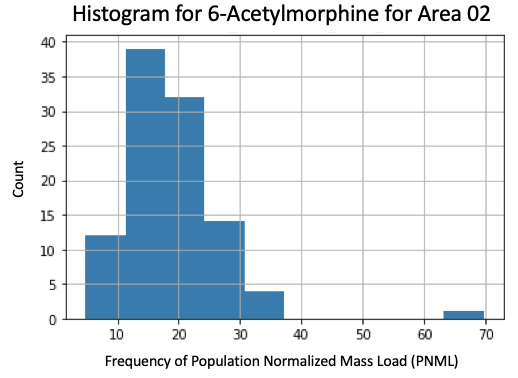
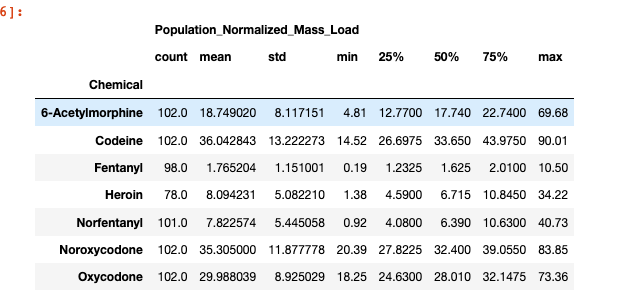
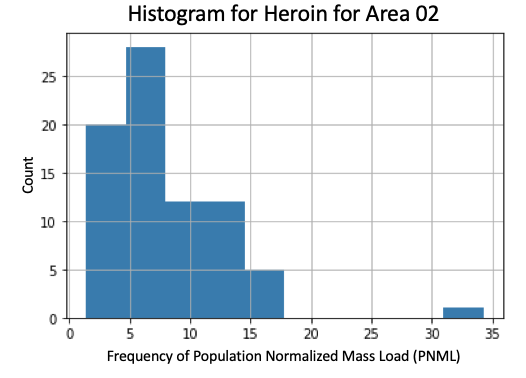
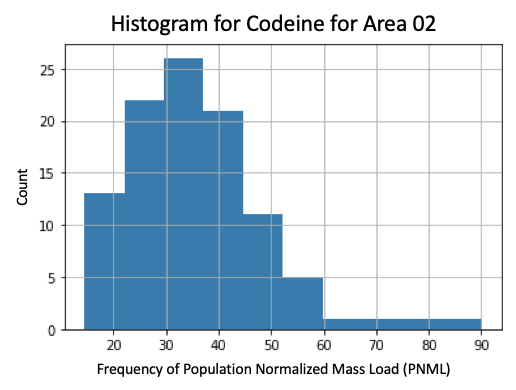
<https://github.com/madelinew/DataRetrieval>

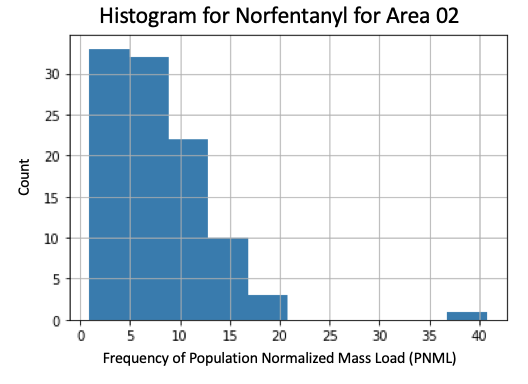
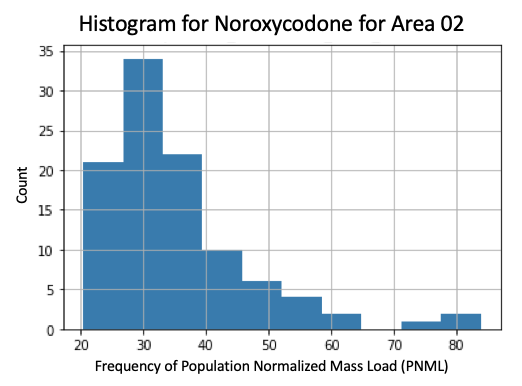
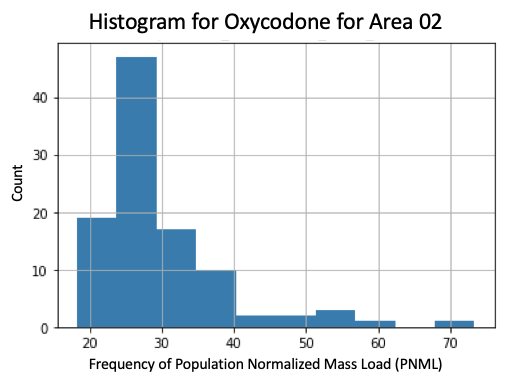
**2. Present descriptive statistics of these data, including a list of samples, features, measures of central tendency (mean, median, or mode, as appropriate) and histograms of each feature**

* There were three collection sites.
* Once the preprocessing was completed, there were four features for each collection site.
  + Site - Area the sample was pulled from.
    - Either TP02, TP04, or TP05
  + Sample\_Date - date that the sample was taken on.
    - Range is from 5/14/18-8/25/19
    - All drugs were tested each day.
  + Chemical - the chemical that was tested
    - Either 6-Acetylmorphine, Codeine, Fentanyl, Heroin, Norfentanyl, Noroxycodone, or Oxycodone
  + Population Normalized Mass Load (mg/day/1000 capita) - is the calculated measurement of drugs that are detected per day for a set number of people within the city (“Fighting Opioid Misuse by Monitoring Community Health”, 2019)

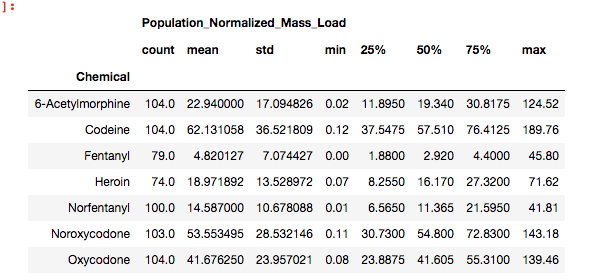
To further clarify how the Population Normalized Mass Load is calculated please refer to the image below (“Fighting Opioid Misuse by Monitoring Community Health”, 2019).

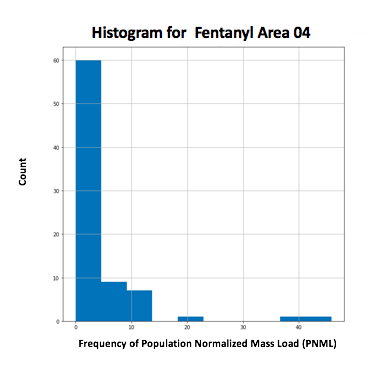
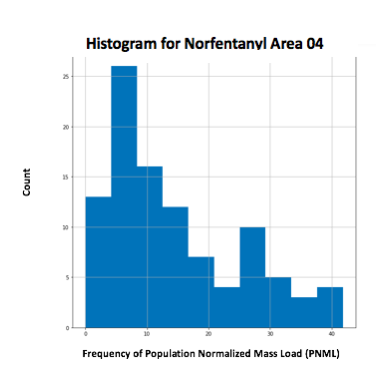
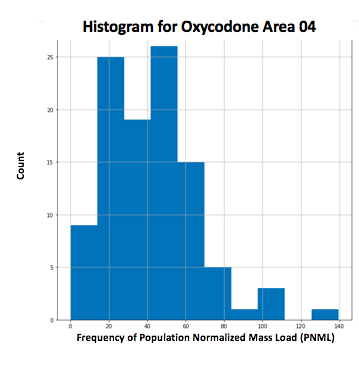
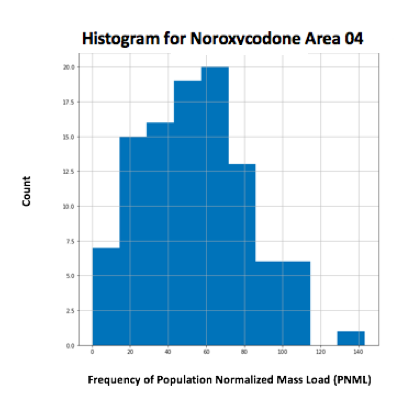
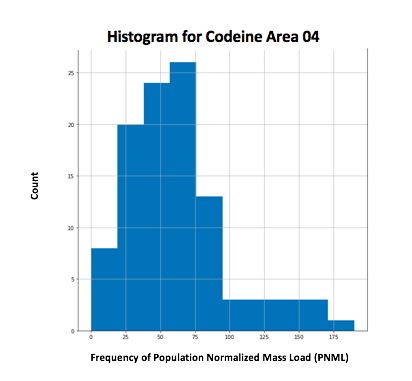
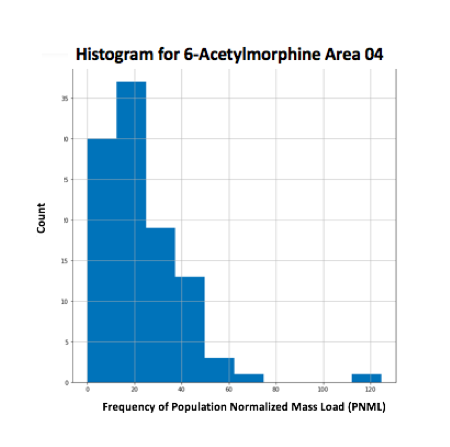
The charts below show the measures of central tendency and histograms for each chemical based on the collection site.

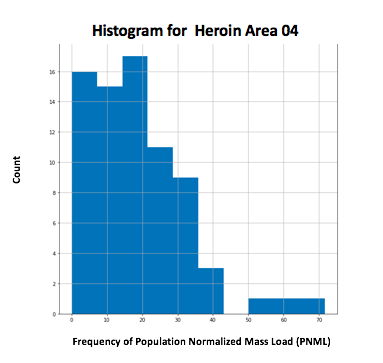
**Collection Site TP02:**



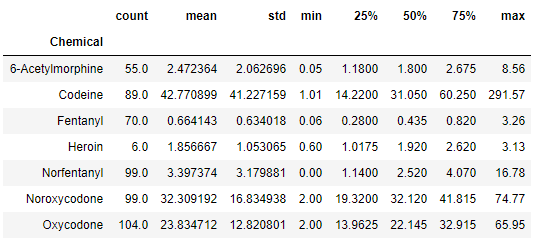
**Collection Site TP04**

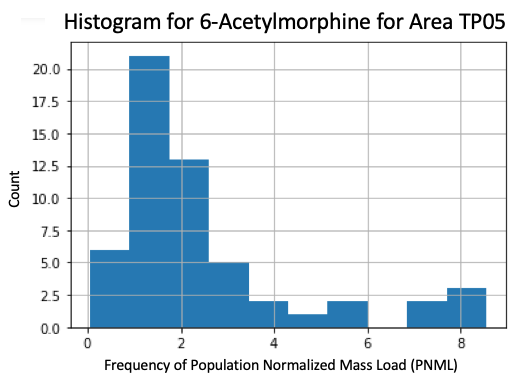
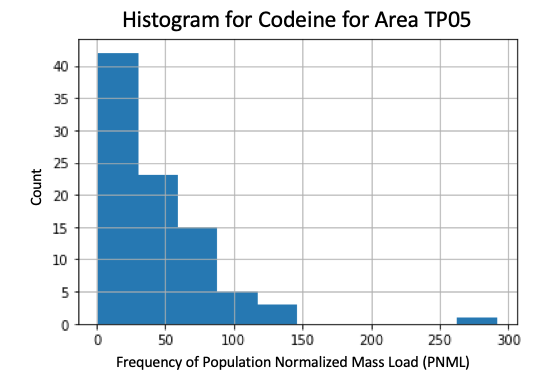
****

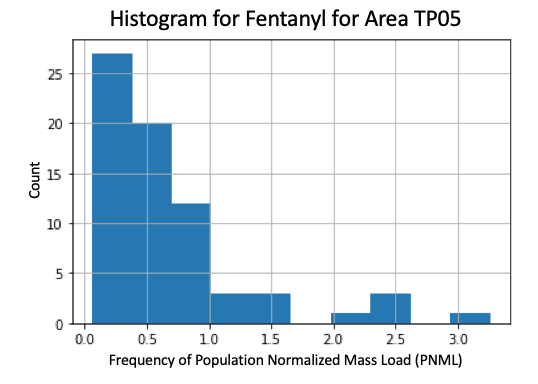
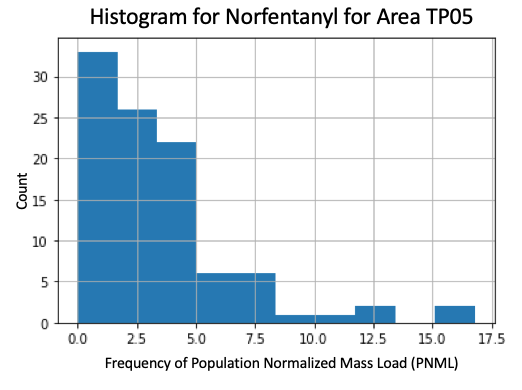
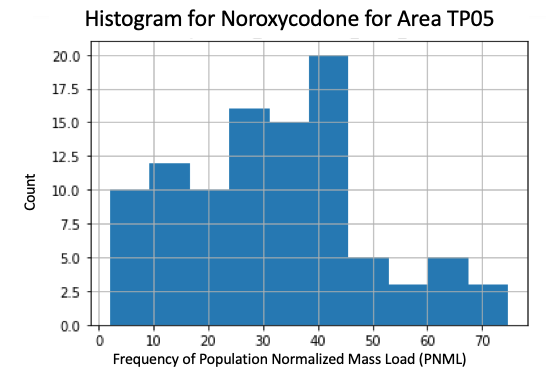
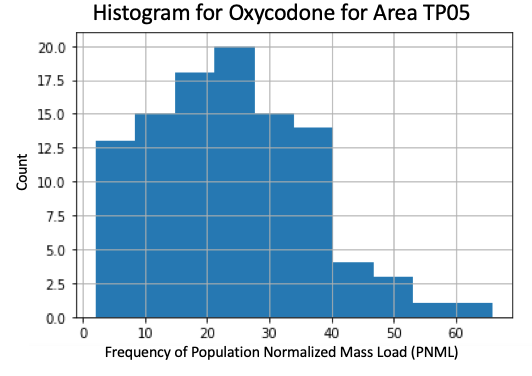
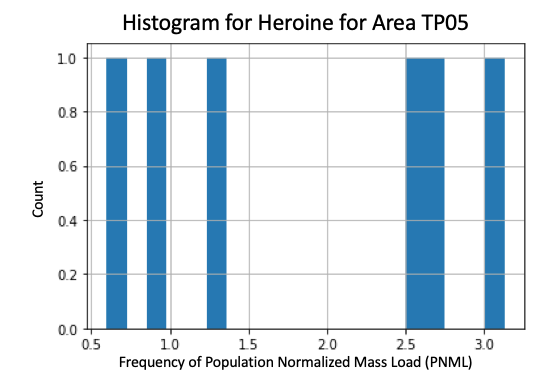




**Collection Site TP05**

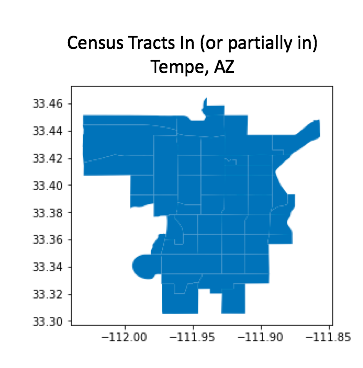
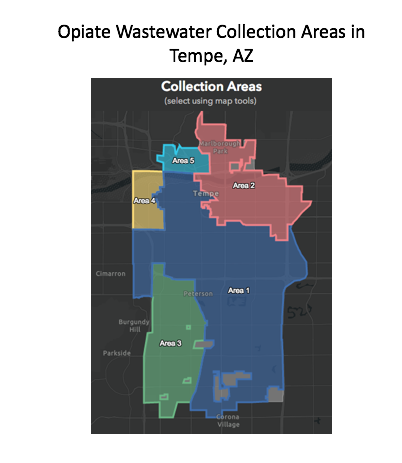






**3. What is your unit of analysis? Why? We discussed this last time, but now rethink it based on your sampling plan.**

Our units of analyses are census tracts within the five Tempe, Arizona collection sites that the Arizona State University Biodesign Institute researchers use to extract the opiate population normalized mass loads. Census tracts provide an overview of demographics, economic, and other social data that can be useful covariates in our analyses. The population normalized mass load values sampled from the collection sites represent an average measure of opiate use per 1000 individuals in each census tract contained in that site. We are assuming that these averages hold constant across each tract in the dataset. This presents some risk of an ecological fallacy that the averages may deviate from tract to tract at the smaller sub unit. However, for now, we are assuming that they generalize well enough for each subunit because the wastewater samples represent a mixed estimate of all individuals within the collection sites.



**4. What is the theoretical population from which this unit will be selected? Why?**

The theoretical population is the entire population of drug users in Tempe Arizona. In an ideal world where this data were accessible, this is what we would want to use to measure the total volume of opiate estimates in the city. We hold our theoretical population just to Tempe, because it is unlikely that estimates of use will be very externally valid compared to other cities, states, and countries due to differences in drug accessibility, regulation, and treatment factors in local, state, and national levels.

**5. What is the accessible population from which this unit will be selected? Why?**

The accessible population is the drug users who live in the 3 collection sites worth of data that we are able to observe in Tempe. These observations represent the drug users we are able to detect and study.

**6. What is the sampling frame that will you use to select your sample? Why?**

The sampling frame is the drug use per 1000 estimates derived from the normalized mass load calculation. These represent our best estimates of drug use within the sampling area based on what we are able to observe using the wastewater data.

**7. How large will your sample be? Why?**

So far we have a dataset which has around 2400 observations (made up of three collection sites, with each site containing around 800 observations). Because the study is still being conducted, this dataset will expand overtime.

**8. What are some threats to the external validity of your sampling Strategy?**

Our study is not likely to be very externally valid for cities outside of Arizona. A main reason for this is that different localities tend to have different regulatory, public health, and criminal dispositions towards drug use that affect the climate in which users can access the drug. For example, policing and enforcement attitudes set by local police departments can affect the incentive structure of drug users and their behavior within the sample. Conversely, cities that prioritize treatment programs over criminal penalty may also influence different behaviors. These same differences exist at the state level as well where differences in regulation and who is authorized to prescribe opioids create differences in accessibility. For example, Arizona allows Opioid distribution at common pharmacies like CVS or Walgreens, while other states do not. This creates the potential for confounding relationships when generalizing to other localities.

Because of these idiosyncrasies, it is risky to assume that Tempe is necessarily representative of other cities in Arizona or the nation. As such, we will present our findings as suggestions for other researchers to look at when they construct their own models.

**9-10. Will you use probability sampling or non probability sampling? Why? What specific sampling strategy will you use? Justify this by relating it back to your research question.**

We are using a non-probability convenience sample of all of the data available. This is appropriate because of the limited sample size observable in the city at this point in time. While it is possible to conduct a random sample of the census tracts within each collection site, we believe it would be inappropriate to do so because it would drastically reduce our sample size and amplify concerns of the ecological fallacy.

We believe it is appropriate to include information on each observation in the data at the tract level so that we can include features of the sample population that may help explain (or more likely just relate to) the estimates of drug use per 1000 in the data.