Data 102 Final Project

Alexus Lopez, Xinqi (Tony) Yu, Lauren Tang, Jahsun Barrack

Data Overview

The data chosen is a subset of the CDC's Annual State-Level National Chronic Disease Indicators (CDI) public dataset. The development of this research included the census of United States hospitalizations and deaths as provided by state hospital discharge data from AHRQ, death certificate data from vital statistics agencies, and population estimates from the U.S. Census Bureau or suitable alternatives, as reported by the CDC.

There are no known groups that were systematically excluded from the study. Selection bias may be a concern if states are better or worse at providing the relevant data. It is unknown whether the agencies involved in data collection specifically notified their patients of data collection; however, the data has been anonymized by providing aggregates on the state level. Interpretable data is provided as annual hospitalization and death rates due to asthma among the general hospitalization and death rates per state. This may affect the interpretation of our findings by lack of more detailed patient information to analyze the impact of all possible confounders. The CDC website includes some important comments on the limitations of the data as follows:

> "The use of a population based-measure can be misleading as it is affected by changes in prevalence over space or time. As one person can have multiple hospitalizations for asthma in a single calendar year, this indicator describes rate of events, not rate of persons hospitalized."
> "The reliability of death certificate data for asthma has been questioned, particularly for older age groups… leading to an overestimation of asthma deaths among people age 55 and older." [1]

As the data is from a national census and we are not analyzing asthma prevalence in a larger population beyond recorded hospitalizations and deaths, selection bias and convenience sampling were not of concern in our study.

Considering the limitations outlined by the CDC above, we believe more specific information about mortality and hospitalization rate for different age groups would allow for more accurate analysis of both of our questions. Additionally, it may be useful to have more specific (e.g. city-based) location data, since air quality likely varies between cities. Different stratifications of gender and ethnicity per instance would have been useful to analyze in tandem.

Research Questions

Our questions were as follows:

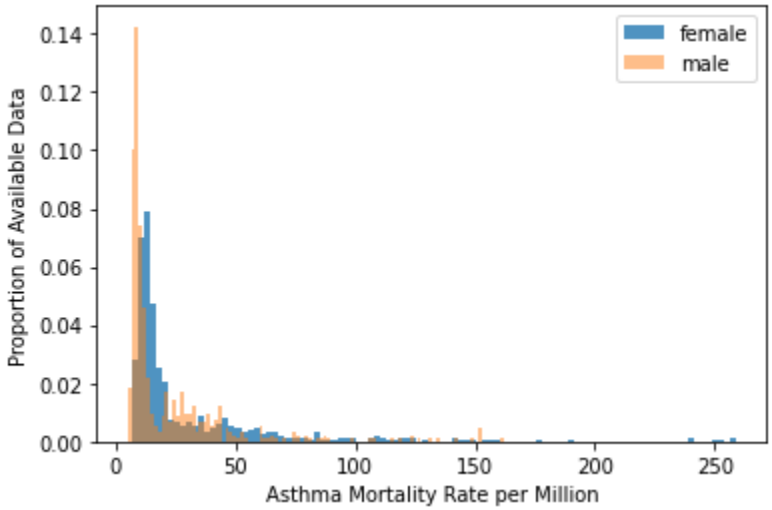
- Is there a relationship between geographic location and hospitalizations for asthma?
    - For states or regions with higher numbers of asthma hospitalizations, public health interventions could be made, such as investigating environmental safety, determining causes of elevated asthma risk, and developing improvements.

[1] https://www.cdc.gov/cdi/definitions/asthma.html

    - Multiple hypothesis testing is a good fit for this question for two reasons. First, we discovered from our EDA that some states have a much higher number of asthma hospitalizations than other states. We suspect that this is not simply an outlier, but an indicator of a larger relationship between geographic location and asthma hospitalizations. Hypothesis testing can test this idea against a null hypothesis stating the difference is simply due to chance, and determine its statistical significance. Second, we would like to compare the hospitalizations for asthma for specific pairs of geographic locations, and since we have many of these hypotheses that we are interested in testing (one for each pair) multiple hypothesis testing with proper corrections in particular is a good fit.
- Is there a causal relationship between gender and asthma mortality rate?
    - If there is indeed a causal relationship, further real-world research into corrective measures for this phenomenon should be taken.
    - Causal inference is a good fit for this question because there are potential confounders that affect both the treatment (gender) and the outcome (asthma mortality rate). Another method like hypothesis testing would only be able to determine if there is a relationship between the two variables, but causal inference can more robustly isolate the effects on the outcome from only the treatment.

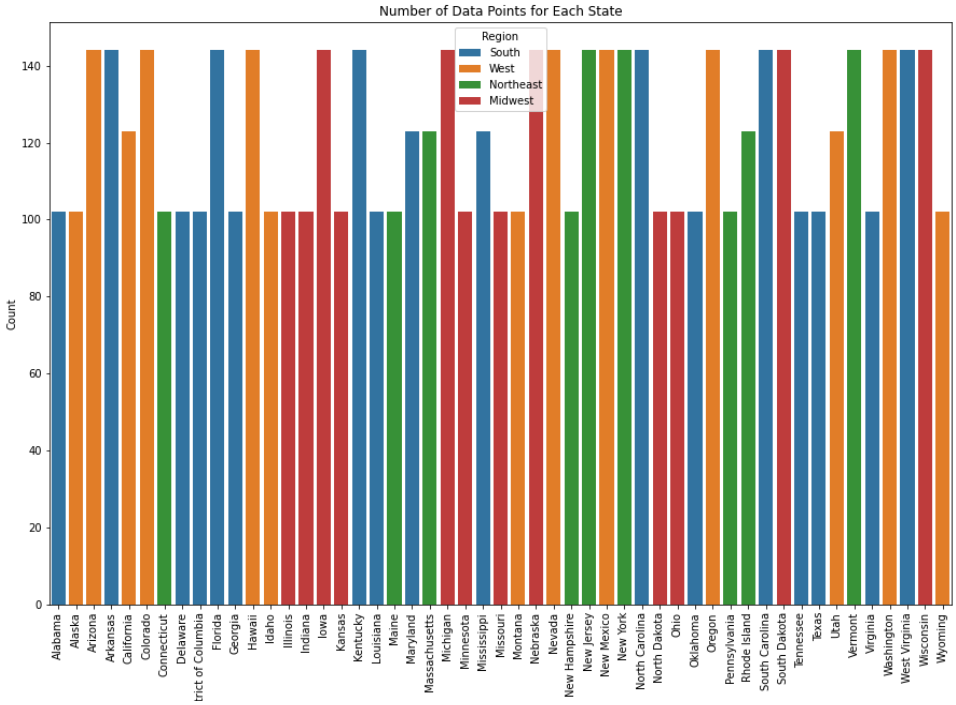
EDA

Since the target data was a subset of a very large dataset, many categories were completely full of null values. These columns were ignored entirely. To explore a potential relationship between gender and asthma mortality rate, the data was limited to instances with "Asthma mortality rate" in the Response column. The data value corresponds to the mortality rate per 1,000,000 people. Two subsets solely comprised of the "Female" and "Male" values for "Stratification1" were created for comparison. As is evident in the visualization below, there is a slight difference in the mean asthma mortality rate per million between the male and female subsets, which indicates that gender may deeply impact asthma mortality rate and even have some causal relationship.



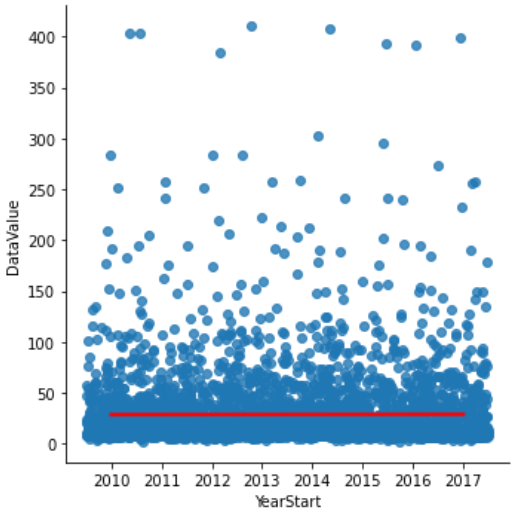
It is important to note that out of 9816 rows, only 2448 rows were focused on female or male mortality rate; the rest were stratified by ethnicity rather than gender. This may affect the accuracy of our conclusions as it significantly reduces the potential values we can analyze.

Representation of Each State in the Data



This visualization shows that the number of data points for each state and region is roughly uniform. While some states have more or fewer data points than others, there does not seem to be a pattern among the different regions. This suggests that there are no systematic biases present in the data such as underreporting or having a sample not representative of the population that would affect our hypothesis testing, so we can more confidently move forward with our tests. We opted to not include instances with the location labels “US” because it provides no specific location information, but this shouldn’t significantly change our results because there were relatively few such instances.

YearStart vs. DataValue (Asthma Mortality Rate)

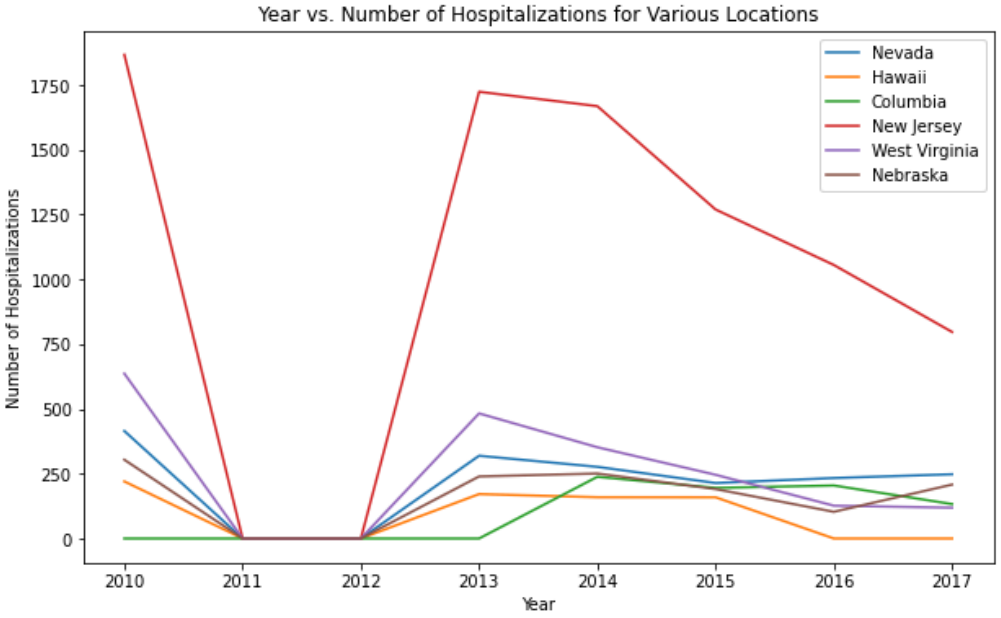


The nearly-flat regression line indicates a surprisingly weakly positive relationship between YearStart and DataValue. We may want to follow up on this because we expected that in recent years, asthma mortality rates would decrease due to medical advances such as better treatments. It would be interesting to determine if asthma treatment progress has been stalling recently.

While plotting, we noticed that there are a few outliers for DataValue (values above 3000), which were removed. This may mean that our results won't generalize well to abnormally high asthma mortality rates. Lastly, since YearStart is not a continuous variable, we added some random noise to it to make the plot more meaningful, drawn from the distribution Normal(0, 0.5). Since the noise is random, it should not impact our model and inferences.

This visualization is relevant to our research question in two ways. First, it suggests that year is likely not a potential answer to our question because of its weak relationship with asthma mortality rate. Second, it motivates questions about this lack of negative relationship: for most diseases, the outlook often improves with time due to medical advances, but this doesn't seem to be the case for asthma as of late, which raises questions as to why, and whether this trend was present before 2010 (the earliest year available in our dataset).

Quantitative Visualization of Location vs. Hospitalizations



There weren't many quantitative variables, so we visualized the year versus the number of hospitalizations, stratified by state, in case the year is a confounding variable that could affect our hypothesis testing. We picked a representative state for each major region of the United States (West, Midwest, South, and Northeast), including Hawaii and the District of Columbia for added geographical diversity. We then selected states that had data collected for the most years within each region, although no data was collected for ANY state during 2011 and 2012. We grouped by year after filtering on the state and averaged the number of hospitalizations.

The most obvious phenomenon is the drastically higher yearly average that New Jersey has in comparison with the other states. Hawaii, by contrast, has a consistently low average. We can't

determine if this is mainly due to asthma incidence levels being affected by location, or if more data was collected in New Jersey than in Hawaii. Additionally, there seems to be a general decrease in hospitalizations over the years, although this could be due to a decline in reporting since the study spans 8 years. We are curious to see if New Jersey's high amount of hospitalizations is simply an outlier or indicative of the very thing we're attempting to test -- that geographic location does in fact influence the number of hospitalizations for asthma.

Question 1: Causal Inference

Methods

The treatment variable is gender with female corresponding to treatment = 0 and male corresponding to treatment = 1.

We identified several confounders.

1. Location is a confounder because the male to female ratio is not 1:1 for all the states.[2] Therefore, it is probable that factors that vary across location such as access to healthcare, lifestyle, and environment would confound the outcome.
2. Race is also another potential confounder. It is likely correlated with income, access to healthcare, and other medical conditions which may all influence asthma mortality.
3. Year is a third confounder as the asthma mortality rate fluctuates year over year for many reasons.[3] We therefore need to compare outcomes for men and women in the same year.
4. Age is the final confounder we identified. Unsurprisingly, older individuals are more likely to die from asthma than younger individuals, so it needs to be adjusted for.

We think that the unconfoundedness assumption holds true for these confounders, because the potential outcomes -- (not) dying due to asthma -- are the same for any individuals within each confound stratification. No individual is entirely immune to developing and dying of asthma.[4]

To handle the identified confounders, we used matching. The female asthma mortality rate for a state in a given year was matched with the male asthma mortality rate for the same state in the same year. The dataset provided the mortality rates as age-adjusted rates, so we didn't have to handle age as a confounder ourselves. The dataset stratifies on either gender or race and not both at the same time, but because the distribution of races within a state is likely equal between men and women, matching by state effectively adjusts for race.

Results

After conditioning on location and year, the only available separate features, the ATE we calculated from the matching method discussed in lecture was -11.0. This means our analysis found that females are 11% more likely than males to die from asthma solely due to their sex, if the unconfoundedness assumption holds..

---

[2] https://worldpopulationreview.com/state-rankings/male-to-female-ratio-by-state
[3] https://www.cdc.gov/cdi/definitions/asthma.html
[4] https://www.webmd.com/asthma/guide/adult-onset-asthma

There is uncertainty in our results because we were only able to use 193 pairs of male and female mortality rates, which is not that large of a sample size. Additionally, there may be some bias introduced by state-year combinations, since not all states had data for the same years.

Discussion

There were several limitations to our methods. First, the dataset did not include multiple stratifications for each instance, as we would have liked. This meant that we were not able to compare the mortality rate between white females and white males for example, only males and females of all races pooled together. Additionally, the dataset became severely restricted when attempting to apply our analyses. Access to more state data could alter our results.

Question 2: Hypothesis Testing

Methods

The hypotheses we tested to answer this are as follows (in regards to U.S. regions):

- There will be more hospitalizations in the Northeast than in the South annually.
- There will be more hospitalizations in the Northeast than the West annually.
- There will be more hospitalizations in the Midwest than the South annually.
- There will be more hospitalizations in the Midwest than the West annually.
- There will be more hospitalizations in the South than the West annually.

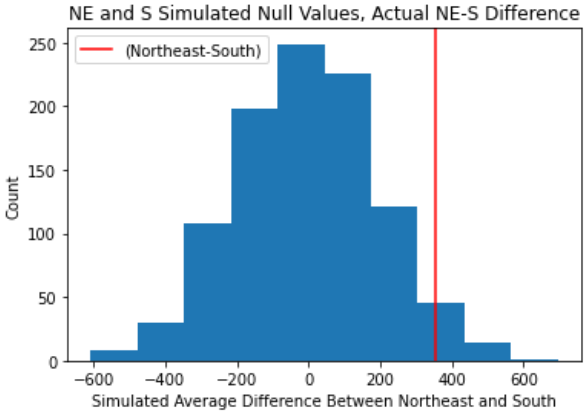
It makes sense to test many hypotheses instead of just one because there are many categories of geographic locations, and we would like to compare the hospitalizations for asthma for specific pairs of them and determine if any geographic location has abnormally high or low hospitalizations compared to another. For example, one specific pair of geographic locations of interest is Southern states versus Midwest states, since different climates, cultures, and other characteristics of the two regions may affect the hospitalization amounts. Moreover, we can get more specific detailed information to see if a significant difference is evident across all of our hypotheses, which contributes to answering the more general question.

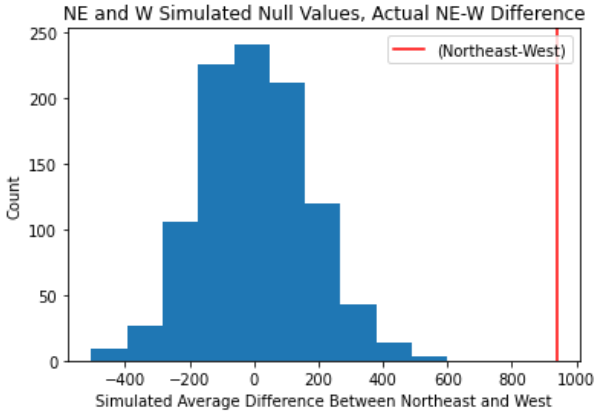
We used permutation testing. The data for region 1 and region 2 were combined. Then, the number of hospitalizations for each instance were shuffled, allowing us to simulate the null hypothesis that there aren't regional differences in the number of hospitalizations. We then bootstrapped samples for each region to obtain an approximate distribution of the difference in average number of hospitalizations, and obtained a p-value for this difference.

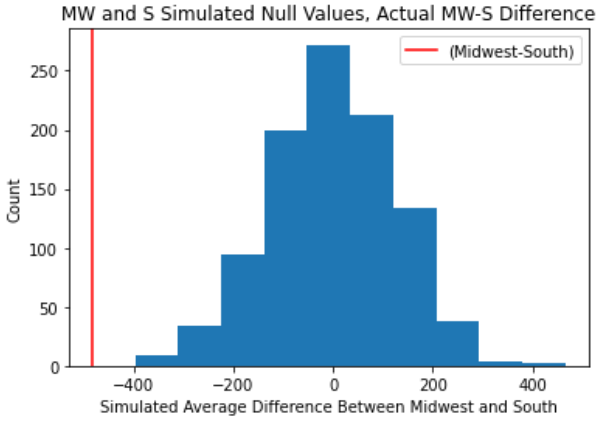
There are two different ways that we corrected for multiple hypothesis tests: the Bonferroni and Benjamini-Hochberg corrections. For the former, we tested n=5 hypotheses with FWER <= alpha=5%, so that each hypothesis was tested with significance alpha/n, or 1%. This controls for the error rate FWER, which is the probability that any of the 5 tests is a false positive. For the latter, we tested n hypotheses with FDR <= alpha. This controls for the FDR, which is the expectation of the FDP over the randomness in data.

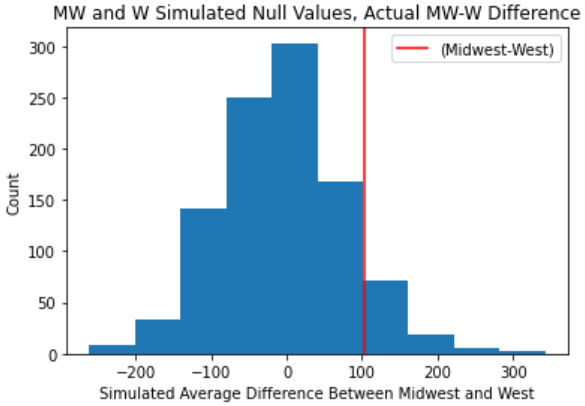
Results

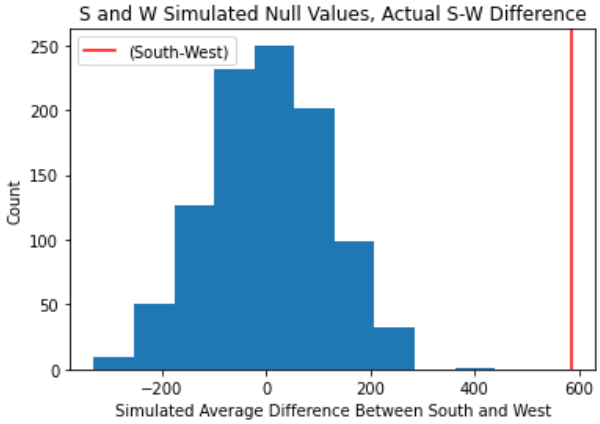
Below are the results from permutation testing for each hypothesis.











Hypothesis	p-value
Northeast-South	0.042
Northeast-West	0.000
Midwest-South	1.000
Midwest-West	0.095
South-West	0.000

After applying the Bonferroni correction, we rejected null hypotheses 2 and 5, whereas we originally rejected 1, 2, and 5. We therefore made 1 false discovery, and our FWER = 0.2. In the

Benjamini-Hochberg correction, just as with the Bonferroni correction, we rejected hypotheses 2 and 5 instead of 1, 2, and 5, and made 1 false discovery.

The Bonferroni correction is more strict than the Benjamini-Hochberg because the threshold for rejection decreases with the number of tests, although both gave the same result here.

Discussion

The results from both error correction methods suggest that we shouldn't reject the first null hypothesis (comparing Northeast and South hospitalizations). Because the results were the same, valuing which metric is most applicable in terms of our concerns over FWER vs. FDP was not a concern. Instead, if we are rejecting the overall null hypothesis based on a majority vote, our original decisions advise rejection, but the corrected ones indicate that we would fail to reject it.

This is where the interpretation of the 3rd null hypothesis (Midwest vs. South) is most relevant. Although a p-value of 1 means the hypothesis itself holds pretty much no value, we could get very different results from simply switching the Midwest and South, because the p-value was not calculated using an absolute value, and the mean of hospitalizations in the Midwest minus the mean of hospitalizations in the South and vice versa imply the same thing -- a difference in hospitalizations due to location. We tested the reverse of this hypothesis and obtained a p-value of 0.0. We would reject this re-formulated hypothesis, even after the error corrections.

With this new p-value and result, we have the proper decisions considering our combination of sub-hypotheses to reject the overall null hypothesis, suggesting that there is in fact a relationship between geographic location and number of hospitalizations for asthma.

P-hacking was avoided by nature of the wide context of census data and establishing statistical methods of assessment that were not manipulated following application. The limitations of our particular study stemmed from the very small number of hypotheses tested, in addition to the implications of our inability to isolate and compare the effects of confounders. Additional hypotheses should be developed and tested to ascertain a more credible result, and previously identified potential confounders should be analyzed.

Conclusions

Given the significant ATE value resulting from our causal inference analysis, if our unconfoundedness assumption holds as a result of conditioning for as many of the relevant confounders as possible, we can tentatively conclude that gender does in fact have a causal relationship with annual asthma mortality rates. This conclusion stems from national data, meaning interventions could be applied on the national scale; for example, females afflicted with asthma are more likely to die than men each year in the United States, so we'd advise public health commissions to educate females diagnosed with asthma of their elevated risk.

As for the results of our multiple hypothesis testing, the evidence supported by permutation testing and error analysis expresses that geographical location on the state level has a significant impact on asthma-related hospitalizations. Action should be taken to analyze and improve

environmental safety in locations with elevated amounts of asthma-related hospitalizations, such as the Northeast United States.

As explained in the data overview and discussion of our results, future studies would benefit from more data in these specific areas of study as well as specific stratifications for each instance (ethnicity, age) being separated into different columns to be analyzed in tandem. Age remains an important confounding variable to our research, and could alter the significance of our findings. More data should be gathered, and tests conducted, to increase the credibility of these results.