00 Sustainability Trends Among Disadvantaged Communities

Exploring the Climate and Economic Justice Screening Tool (CEJST) Data

Stefani Langehennig Zach del Rosario

Before starting the activity, please use this QR code to access a survey. Even if you choose not to participate in the associated study, you are still welcome to do the activity!

Please do not move on until your instructor tells you to.

Overview

As climate change continues to impact the world in which we live, numerous initiatives have been started to better understand the influence it has on individuals and communities. One of those initiatives stems directly from an Executive Order (EO) issued by President Joe Biden in January 2021. The EO resulted in the Council on Environmental Quality creating a tool by which the public can track various burdens across a number of communities. The primary aim of the tool is identify and subsequently help communities disadvantaged by these burdens in government social programs.

The Climate and Economic Justice Screening Tool (CEJST) is the result of the EO. While the tool established by the Council on Environmental Quality covers a number of burdens (health, transportation, and workforce development, for example), this activity will focus on the sustainability aspects of the tool, including climate change, energy, and legacy pollution burdens on communities.

To set the stage for this activity, we are going to use the CEJST data to explore whether there is a relationship between the energy burden percentile in census tracts and the share of population of Blacks or African-Americans in census tracts. Below, we will explore the dataset, as well as our variables of interest, in-depth.

Dataset

The data used for this analysis comes from the CEJST website. The columns (variables) we are most interested in for better understanding these data are:

- The *energy burden percentile*, which captures the percentile of energy cost as well as energy-related pollution within a census tract.
- The *percent of African-American or Black alone*, which captures the percent of African-American or Black individuals in a census tract.

First, we will load our data and clean up the variable names using the various packages available to us in the `tidyverse`.

```
# Import tidyverse
library(tidyverse)

# Load CEJST data and create a new dataframe called 'df_raw'
filename <- "../data/1.0-communities.csv"
df_raw <- read_csv(filename)
```

```
# Create a new dataframe called 'df_data' with new column names
df_data <-
  df_raw %>%
  janitor::clean_names()
```

```
# Select only those columns we'll use in this activity
df_data %>%
  select(
    census_tract_2010_id,
    percent_black_or_african_american_alone,
    energy_burden_percentile
  ) %>%
  head(3)
```

```
# A tibble: 3 x 3
  census_tract_2010_id percent_black_or_african_america~1 energy_burden_percen~2
  <chr>                                             <dbl>                  <dbl>
1 01001020100                                        0.07                     49
2 01001020200                                        0.57                      6
3 01001020300                                        0.24                     68
# i abbreviated names: 1: percent_black_or_african_american_alone,
#   2: energy_burden_percentile
```

We will focus on a few columns in this dataset:

- `census_tract_2010_id`: Each row in this dataset corresponds to a *census tract*; this is a small geographic region of the U.S. chosen to represent a consistent number of persons. Census tracts contain about 4000 people, though may contain as few as 1200 and as many as 8000 people.
- `percent_black_or_african_american_alone`: This reports the percent of people in the census tract who are Black or African American.
- `energy_burden_percentile`: This reports the **percentile** of energy burden for each census tract. Energy burden is computed as the average annual cost of energy divided by the average household income within the census tract—this is a measure of how "burdened" a household is by energy expenses. A larger energy burden means a household needs to spend more of its income on energy bills.
  - The percentile is then computed as the *ordering* of energy burden values for each census tract: The 0th percentile corresponds to the smallest value of energy burden, while the 100th percentile corresponds to the largest value of energy burden.

Next, let's conduct EDA to better understand our variables of interest, `energy_burden_percentile` and `percent_black_or_african_american_alone`.

Exploratory Data Analysis (EDA)

To begin, let’s subset our data so we can get our descriptive statistics for our variables of interest.

```
# Subset full dataframe
df_small <-
  df_data %>%
  select(energy_burden_percentile, percent_black_or_african_american_alone)

# Take a look at the subset of data
glimpse(df_small)
```

```
Rows: 74,134
Columns: 2
$ energy_burden_percentile                <dbl> 49, 6, 68, 63, 38, 57, 72, 46,~
$ percent_black_or_african_american_alone <dbl> 0.07, 0.57, 0.24, 0.05, 0.18, ~
```

Let’s use this subset to get our measures of spread and central tendency.

```
summary(df_small)
```

```
 energy_burden_percentile percent_black_or_african_american_alone
 Min.   :  0.00           Min.   :0.0000
 1st Qu.: 24.00           1st Qu.:0.0100
 Median : 49.00           Median :0.0400
 Mean   : 49.55           Mean   :0.1341
 3rd Qu.: 75.00           3rd Qu.:0.1500
 Max.   :100.00           Max.   :1.0000
 NA's   :1054             NA's   :745
```

> Synthesizing Descriptive Statistics
>
> What are some take-aways from our descriptive statistics for our variables of interest? Anything concerning or interesting?

Next, we can plot the percent of Black or African-Americans against the energy burden percentile to see if there is a negative or positive trend between the two variables.

As the figure shows, there is a positive relationship between the percent of Black or African-Americans living in a census tract and the energy burden percentile in the respective census tract.

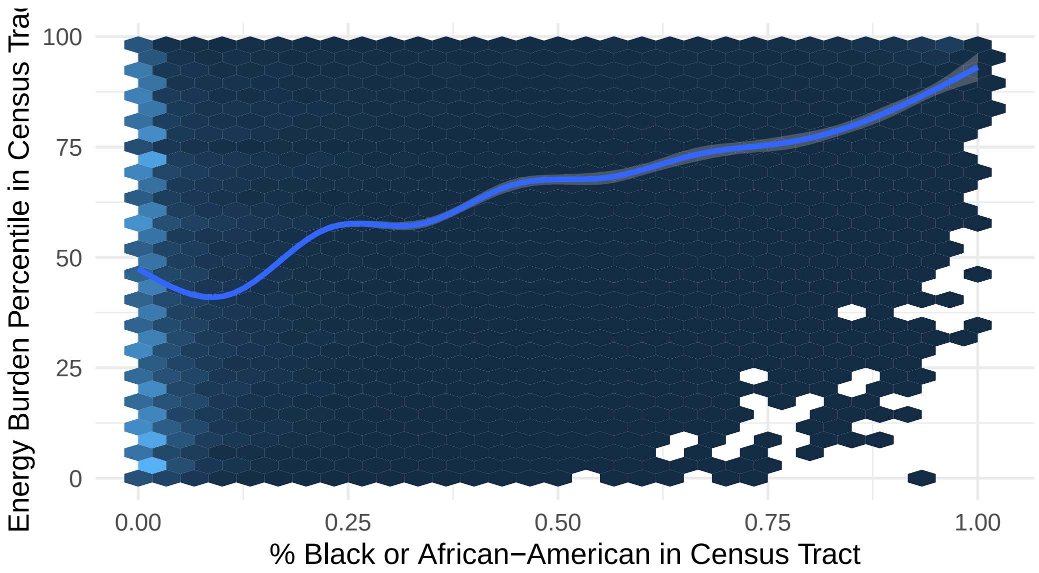


Figure 1: Basic scatterplot of the CEJST data.

> Visualizing Variables of Interest
>
> What other ways might we choose to visualize our variables of interest? Is there anything else concerning or interesting based on our visualizations?

Next Steps

We now have a better understanding of sustainability trends among disadvantaged communities using just a few of the variables in the CEJST dataset. Specifically, we explored the relationship between the percent of Black or African-Americans living in a census tract and the energy burden percentile in the census tract. Our high-level exploratory data analysis uncovered a positive relationship: Black or African-Americans that live in a census tract appear to experience a higher energy burden.

> Other Trends
>
> How might the trend differ by state? Why do you think they will or will not differ from the overall trend observed above?

With this information in hand, we will use these data, as well as other variables in the CEJST dataset, to formulate hypotheses. We will test our hypotheses in both a frequentist and Bayesian framework, comparing the application of both approaches across the different stages of our analysis with the end goal being general inference.

Coming up next: When the instructor tells you, please start working through `01 Introduction to Different Statistical Paradigms` with the other students at your table. **Please do not move on until your instructor tells you to.**

01 Introduction to Different Statistical Paradigms

Comparing Frequentist and Bayesian Approaches

Stefani Langehennig Zach del Rosario

Introduction

This activity is all about *statistical inference* and *statistical assumptions*. Both are important when you have data and want to use that data to help you better understand some phenomenon about the world. For example, in 00 Sustainability Trends Among Disadvantaged Communities, we used the CEJST dataset to understand general trends around sustainability and disadvantaged communities. However, at this point, we have not crafted specific research questions or hypotheses that we would like to rigorously test with the data we have. Nor do we know, with any degree of certainty, how "correct" the trends are that we saw in our data or all the ways that this trend could have occurred.

There are a few different ways that we can test our hypotheses using methods of statistical inference. Statistical inference allows us to take some data, create a model, and make sense of the complex world around us. Each method has a set of assumptions built into it, which influences the approach you take to answer your research questions, as well as the way you interpret your findings. In the activity that follows, you will have the opportunity to use methods of statistical inference to answer a research question and explain, based on the set of assumptions baked into your approach, your ultimate findings.

Learning Objectives

By the end of the activity, you should be able to:

1. Evaluate multiple hypotheses using inferential statistical results
2. Connect your evaluation of hypotheses with real-world factors
3. State the primary statistical assumptions for Frequentist and Bayesian inference, and understand how they can lead to different conclusions

Before We Begin...

Ask yourself the following questions:

> Warmup Questions
>
> - What do you remember about analyzing a dataset with statistics?
> - What do you know about frequentist and Bayesian statistics? (*Note*: It's okay if you don't know what this means!)

Discuss these questions with the people around you.

The Big Idea: Inference

Statistical inference is drawing conclusions about **data we haven't seen** using **only the data that are available.** As a simple example, if we're interested in the energy burden pattern in Colorado but we only have data from Massachusetts, we should be careful when trying to use the MA data to predict patterns in CO. Crucially, **the assumptions we make for inference will strongly affect the conclusions we make.**

All models are wrong....

All statistical analyses involve making assumptions—we'll practice making and assessing statistical assumptions in this activity.

Crafting Research Questions & Hypotheses

Ahead of using data and statistical inference, we usually create theory-based *research questions* and *hypotheses* that help guide what we expect to find. Both are focused on an aspect of the world that we want to know more about. However, research questions tend to be open-ended, allowing for discussion on what the outcome might be and what might explain it. For example, a research question about sustainability and diverse communities may look like this:

> *Why does the projected flood risk vary among different communities in U.S. Census tracts?*

Hypotheses, on the other hand, tend to be closed-ended, articulating a certain relationship between an outcome and the things that influence that outcome. For example, one of the testable hypotheses stemming from the research question on projected flood risk may be:

> *As the population density of Hispanic communities increases (decreases), the projected flood risk increases (decreases).*

Note that this hypothesis is succinct and testable with the data we have. It is also directly related to our more general research question.

Paradigms in Statistical Inference

There are a few different ways to draw conclusions about questions we have using data. We will focus on two inferential perspectives in this class: *Frequentist* and *Bayesian*. For this activity, we will discuss these differences in the context of **general inference** and **model summaries**.

For general inference, we are concerned with answering questions about a *larger* dataset, while only having access to a *smaller* dataset. We do this by:

1. Translating our question into a mathematical model, which requires assumptions
2. Fitting the model using data
3. Interpreting model results in terms of our question

After we have done the general inference steps above, we must think about how to translate what we have done and what we have found.

- Statistical models are complex objects, so we use *model summaries* to help.
- Statistical models represent the *uncertainty due to limited data.*

Some common summaries we use:

Table 1: Common Model Summaries

Point estimate	Uncertainty	Interval
42	4	(42 +/- 2x4) = (34, 50)

Stay Tuned!

Note that we haven't talked about frequentist or Bayesian statistics yet! We'll get there later in the activity….
You can read more about the statistical inference in 03 One Page Summary.

Next Steps

Given the context we have on sustainability trends among disadvantaged communities using just a few of the variables in the CEJST dataset, we can develop a research question and hypotheses to test using methods of statistical inference:

Research Question:

> *Do Black Americans experience a disproportionate level of energy expenditure?*

Hypothesis:

> *As the population of Black Americans increases (decreases), the level of energy expenditure increases (decreases).*

Coming up next: When you get here, please discuss the reading with your tablemates. When you all feel that you (at least somewhat) understand the reading, please start working through `02 CEJST Activity` with the other students at your table.

We are going to test our hypothesis and answer our research question using the CEJST dataset.

02 CEJST Activity

Armed with all of that background knowledge on the CEJST dataset and on statistical inference, we can proceed with a detailed analysis of the data. In particular, we are interested in assessing the following hypothesis:

> *As the population of Black Americans increases (decreases), the level of energy expenditure increases (decreases).*

In this part of the session you will interpret results from a statistical model fitted to datasets from different U.S. States.

Overview

Throughout this activity, you will be studying a statistical model fitted to data from the CEJST dataset. As a reminder, we are interested in the `Energy Burden Percentile` (higher values correspond to a higher burden) and the `Percent Black Census Tract` (which measures the number of people in a region who are Black).

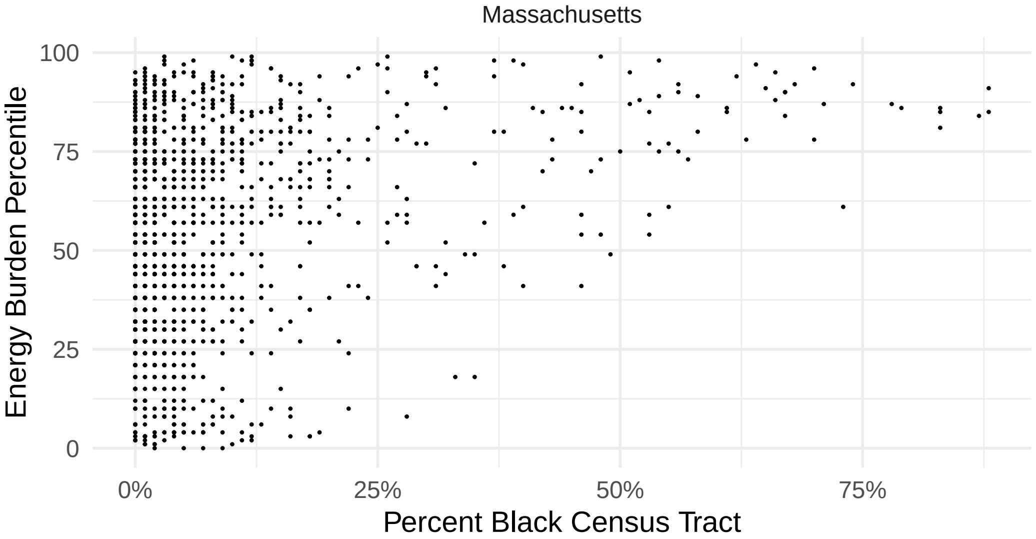


Figure 1: CEJST data for Massachusetts.

From this scatterplot, we can see that the energy burden seems to increase as the percent Black increases. However, we can make this rough observation more formal by using a statistical model.

Analyze a Bayesian Model

To analyze the dataset, we will use the following statistical model,

$$B = mP + b + \epsilon$$

where B is the energy burden percentile, P is the percent Black, m is the slope parameter, b is the intercept parameter, and ϵ is a *residual* term that represents factors not accounted in the model. The residual term is assumed to be normally distributed $\epsilon \sim N(0, \sigma^2)$ with an unknown parameter σ^2. All three parameters have a prior distribution, defined via

$$\begin{aligned} m &\sim N(\mu_m, \sigma_m^2), \\ b &\sim N(\mu_b, \sigma_b^2), \\ \sigma^2 &\sim \text{Exponential}(1/s_y), \end{aligned}$$

where m, b, σ^2 are independent.[1] We will discuss how to set the prior through its parameter values $\mu_m, \mu_b, \sigma_m^2, \sigma_b^2$ later in this activity.

Study the posterior distribution

Results from a Bayesian model take the form of a *posterior distribution* for the model parameters. The model has two parameters that are closely related to our hypothesis: The slope m and intercept b of the fitted line. “Fitting” the Bayesian model to the Massachusetts dataset will result in a posterior distribution for the parameters, with an example posterior given below:

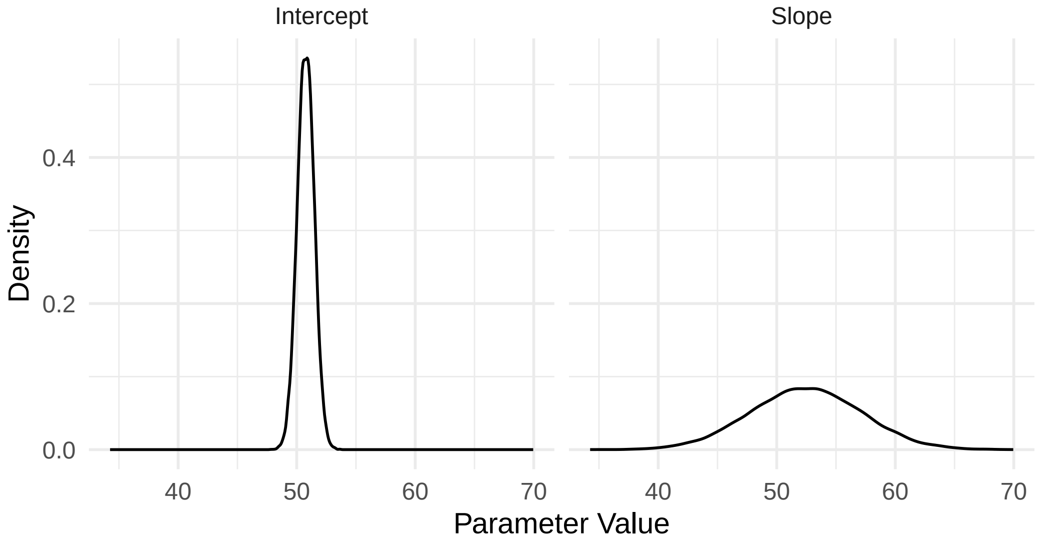
The posterior distribution helps us determine **how confident** we should be in conclusions drawn from the model. The next exercise will help you assess confidence in results based on the fitted model.

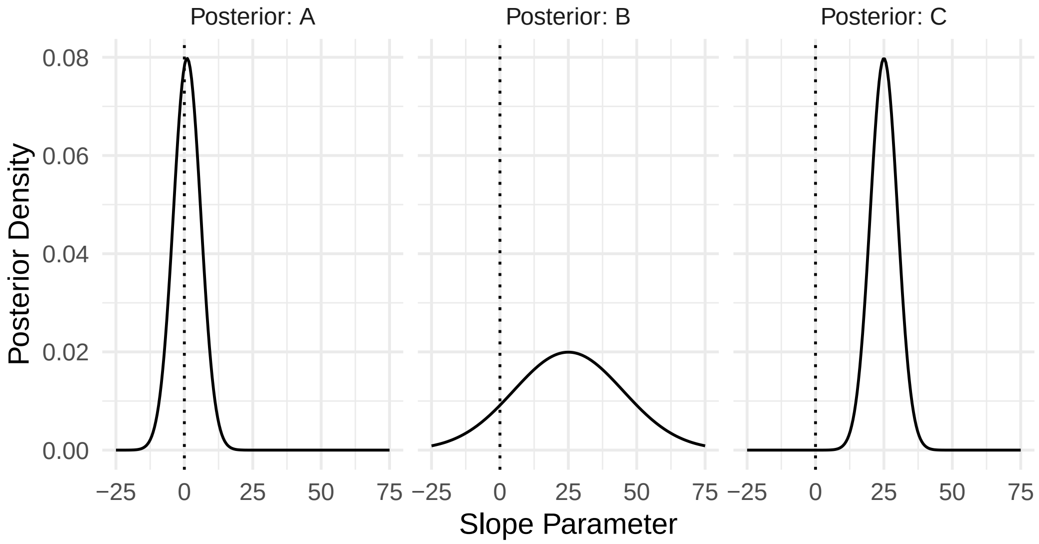
Assessing confidence

Let’s imagine three different posterior distributions for the posterior (marginal) distribution for the slope parameter m.

Model summaries

[1] Note that s_y is determined based on the standard deviation in the observed data. This is a way of *autoscaling* the prior.


Intercept
Slope
Density
0.4
0.2
0.0
40
50
60
70
40
50
60
70
Parameter Value


Posterior: A
Posterior: B
Posterior: C
Posterior Density
0.08
0.06
0.04
0.02
0.00
−25
0
25
50
75
−25
0
25
50
75
−25
0
25
50
75
Slope Parameter

Questions for the Class

- Roughly, what fraction of *Posterior Distribution A* is greater than zero?
  - (Write your response here):
- Roughly, what fraction of *Posterior Distribution B* is greater than zero?
  - (Write your response here):
- Roughly, what fraction of *Posterior Distribution C* is greater than zero?
  - (Write your response here):
- Which posterior gives the *highest confidence* (*highest probability*) that the slope parameter is positive? How can you tell?
  - (Write your response here):

General inference

Questions for the Class

- How does the slope parameter relate to our hypothesis? As a reminder, our hypothesis is:

  > *As the population of Black Americans increases (decreases), the level of energy expenditure increases (decreases).*

```
-   (Write your response here):
```

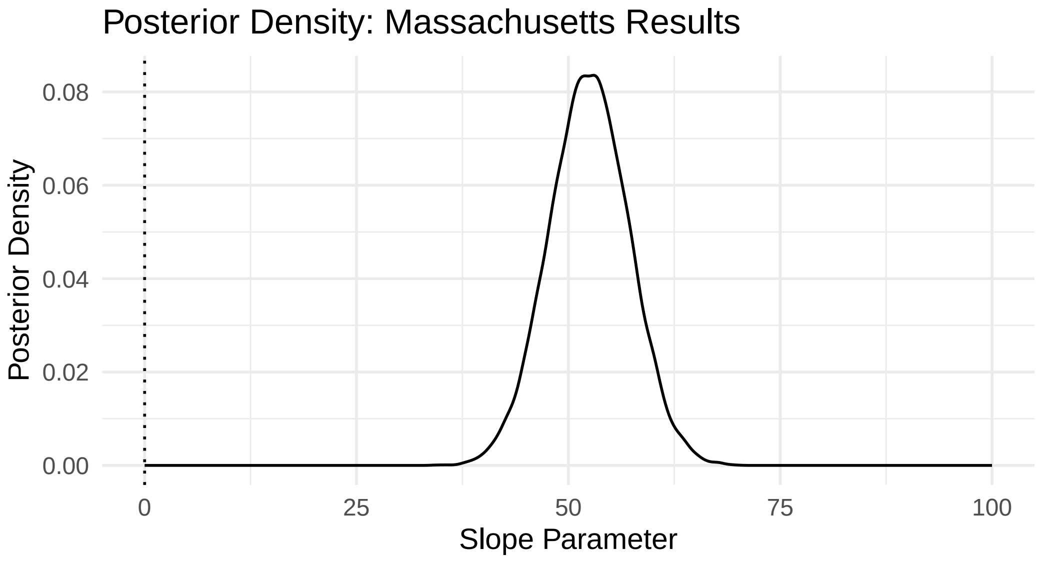
Let's return to the posterior from our model for the Massachusetts data and use the same reasoning as above to make sense of the results.

Questions for the Class

As a reminder, our hypothesis is:

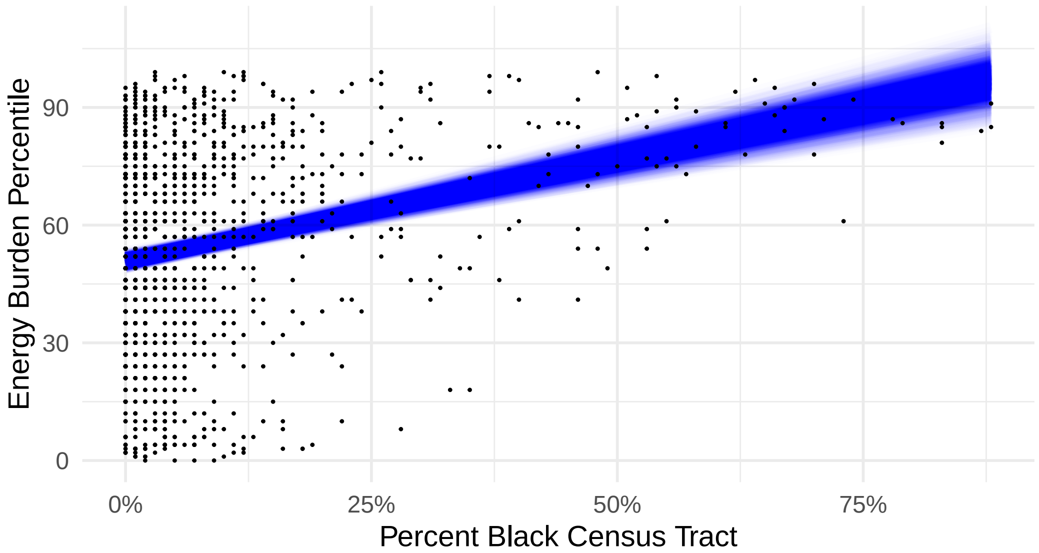
> *As the population of Black Americans increases (decreases), the level of energy expenditure increases (decreases).*

- For Massachusetts, does the fitted model support or contradict our hypothesis?
  - (Write your response here):
- How confident are you in the model results?
  - (Write your response here):



Study the posterior predictions

Bayesian analysis does not produce a "best" line; rather, the posterior distribution implies a *family* of lines (each with a different chance). We call this the *posterior predictive distribution*. For instance, the following visualizes the posterior predictive distribution of lines against the Massachusetts data. This object appears as a "cone" with darker regions corresponding to lines with higher probability (density).



We can use this kind of plot (predictions against observed data) as a way to sanity check the model. You'll do this in the following questions.

Questions for the Class

- Do all data points (black dots) land near the predicted lines (transparent blue lines), or do some dots land far from the lines?
  - (Write your response here):
- This model represents the *overall trend* in the data well. In your own words, describe how the model fits the overall trend in the data.
  - (Write your response here):

The model should fit the data reasonably well; otherwise, we should *distrust* its results. It doesn't matter if the posterior distributions agree with our hypothesis if the model fits the data poorly!

The Prior Distribution

Above, we glossed over how we arrive at a posterior distribution. In addition to the equation for the line we must also provide a *prior distribution* for the model's parameters. In a Bayesian analysis the prior represents all of our *prior knowledge* about the problem.

For instance, if we were confident that the slope of the Energy Burden vs Percent Black line is positive, we could represent that with a prior distribution that was tightly concentrated at a positive value (Case A below). If we were highly uncertain about the slope, we might represent this case with a prior distribution centered at zero with a very large standard deviation (Case B below).

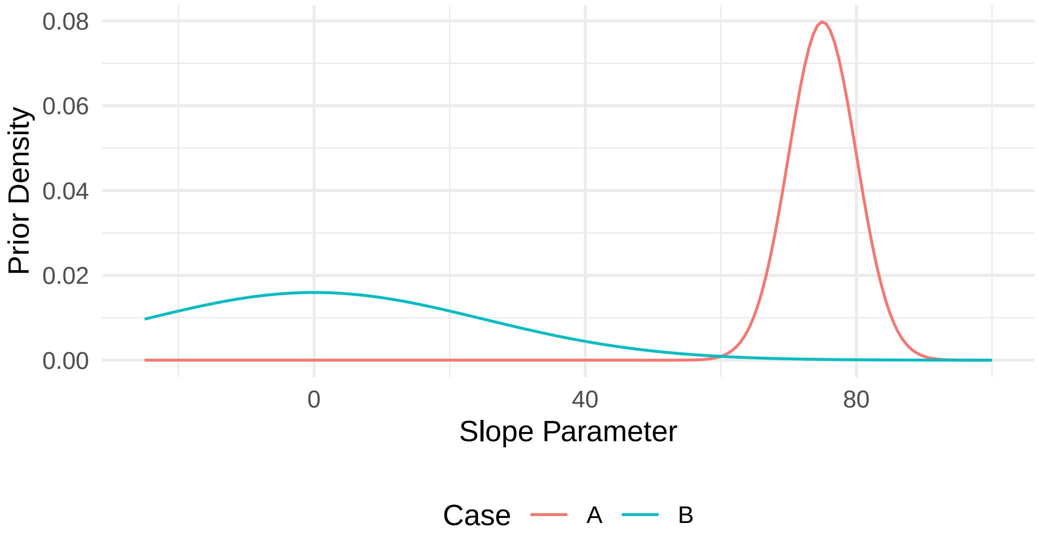
Fitting a model: Data + Prior = Posterior

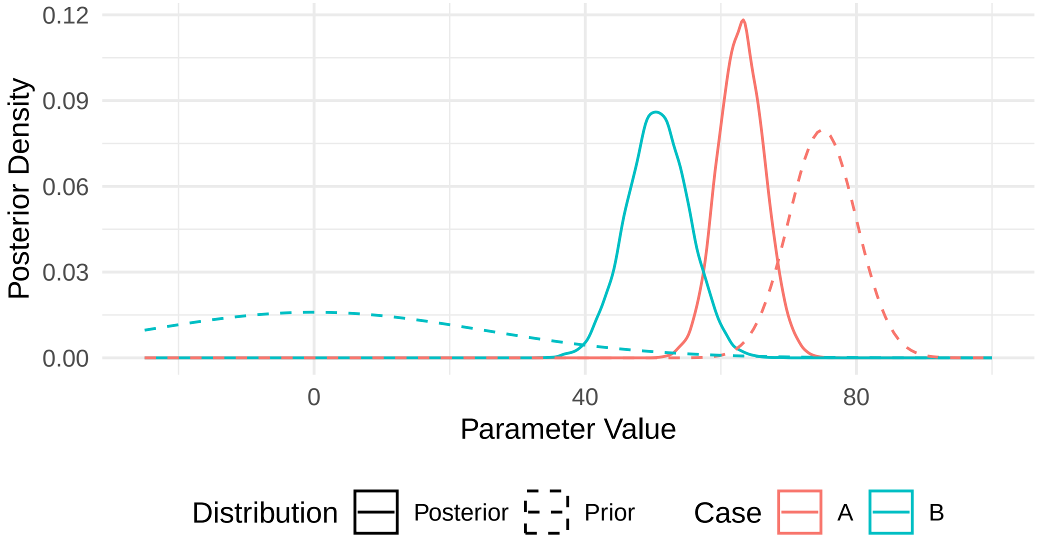
The mathematical details of fitting a Bayesian model are outside the scope of this activity. However, the basic "formula" is:

[Data + Model = Posterior]

Note that the Model contains both the formula for the model $B = mP + b + \epsilon$ and the prior distribution for the parameters m, b, σ^2. With a small dataset the posterior distribution will largely depend on the prior. With a larger dataset, the posterior will depend more on the data.

For instance, the following figure shows what happens when we use two different prior distributions for the slope when fitting the Massachusetts data:

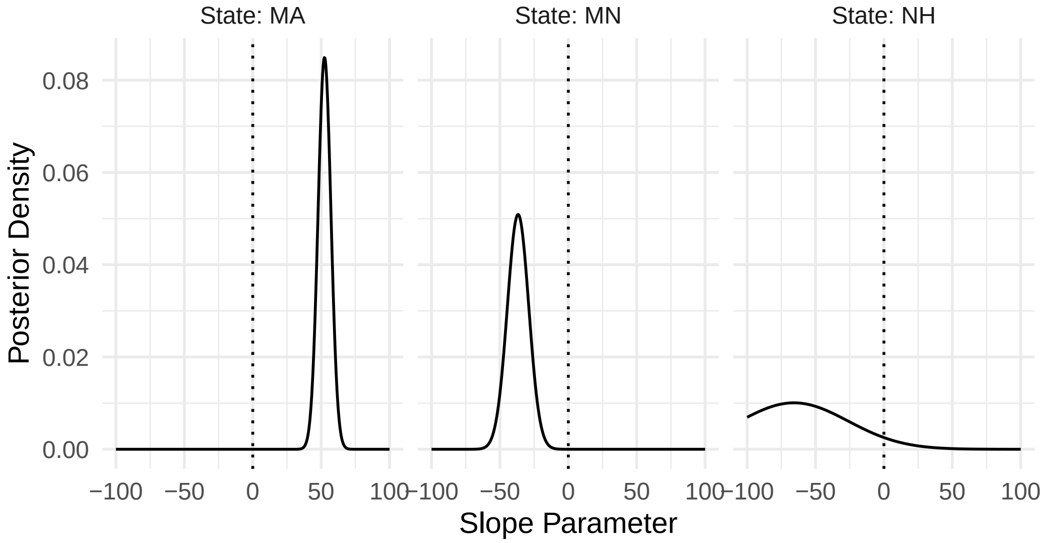

0.08
0.06
0.04
0.02
0.00
Prior Density
0
40
80
Slope Parameter
Case
A
B


0.12
0.09
0.06
0.03
0.00
Posterior Density
0
40
80
Parameter Value
Distribution
Posterior
Prior
Case
A
B

Note that the highly uncertain Case B is shifted quite a bit more (from prior to posterior) than the Case A results. If we were to obtain a larger and larger dataset, using either prior (Case A or B) would converge to the same posterior distribution.

Three options

Bayesian approaches to statistics are particularly useful when we have limited data, as they allow us to incorporate prior knowledge. For the rest of the activity we'll consider a scenario where our access to the CEJST data is limited: Suppose we are conducting our analysis while the data are actively being gathered. In this case, we may have access to the data for some states before others. In this context, we can conduct a *sequential* Bayesian analysis by using the posterior from one analysis as the prior for a new analysis.



Pick a State

Study the posteriors above carefully; you will use this as a *prior distribution* for the slope for the rest of the activity. This means you will combine *new data* with a *prior distribution* to form a new *posterior distribution* for the model parameters. The *prior distribution* should reflect your beliefs about what you think the slope parameter should be.
Pick *one state for your group*, then come ask the instructor for your chosen state's packet.

Armed with this fundamental understanding of statistical inference, we can now apply these ideas to study data from the other states!

Colorado

Study the results

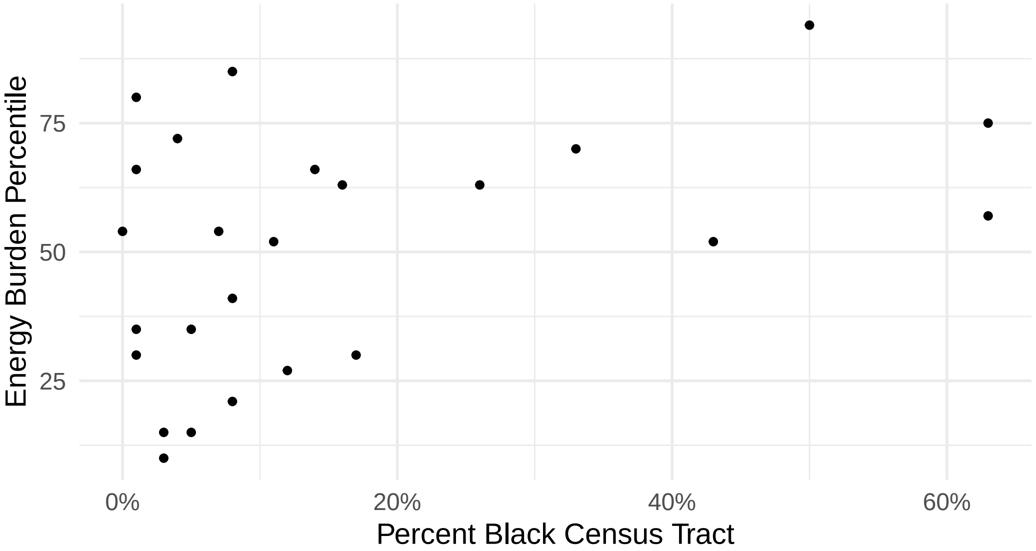
(Look at the packet you got from the instructor to answer the following questions.)

Studying Model Results

- What does the model suggest about the trend of energy burden with percent Black in Colorado?
  - (Write your response here):
- How well do the model predictions match the data?
  - (Write your response here):
- How confident are you in your conclusion?
  - (Write your response here):

Florida

In some cases, we may find that gathering more data is simply not possible. Let's suppose that, for some reason, Florida is unwilling to provide all of their energy burden data. Therefore, we must figure out what to do with only $n = 25$ observations:



Given the limited data, our results will depend much more strongly on our prior distribution.

Study the results

(Look at the packet you got from the instructor to answer the following questions.)

Studying Model Results

- What does the model suggest about the trend of energy burden with percent Black in Florida?
  - (Write your response here):
- How well do the model predictions match the data?
  - (Write your response here):
- How confident are you in your conclusion?
  - (Write your response here):

Full USA

After waiting some time, we finally get access to the full U.S. CEJST dataset. With such a large dataset, we expect to see that the results will not depend so much on our choice of prior distribution.

Study the results

(Look at the packet you got from the instructor to answer the following questions.)

Studying Model Results

- What does the model suggest about the trend of energy burden with percent Black across the whole U.S.?
  - (Write your response here):
- How well do the model predictions match the data?
  - (Write your response here):
- How confident are you in your conclusion?
  - (Write your response here):

Coming up next: When the instructor tells you, please start working through `04 Activity Wrap-up` with the other students at your table. **Please do not move on until your instructor tells you to.**

(This page intentionally left blank)

03 One Page Summary

Stefani Langehennig   Zach del Rosario

Overview

You can use this one-page summary throughout the activity as a reference! You won't need it until `02 CEJST Activity`, though.

General Inference

Inference is all about *cautiously* answering questions about a larger dataset while only having access to a smaller dataset. For instance, we may be interested in national trends, but may only have access to data from a handful of U.S. states.
To perform inference we must:

1. Translate our question into a mathematical form,
2. fit the model using a dataset, then
3. interpret the results in terms of our original question and testable hypothesis.

In order to accomplish (1), we must make a number of *assumptions*. Different statistical approaches use different assumptions....

Model Summaries

Statistical models are complex objects, so we use *model summaries* to help make sense of them. These include:

- Summaries such as (point) estimates
- Intervals, such as confidence intervals

(This page intentionally left blank)

04 Activity Wrap-Up

Comparing Frequentist and Bayesian Approaches

Stefani Langehennig Zach del Rosario

Overview

In this activity, we learned more about *statistical inference* and *statistical assumptions* through the lenses of Frequentists and Bayesians. We used the CEJST dataset to understand general trends around sustainability and disadvantaged communities, as well as crafted specific research questions and hypotheses that we would like to rigorously test with the data we have. Our research question asked:

> *Do Black Americans experience a disproportionate level of energy expenditure?*

And our primary hypothesis we wanted to test is:

> *As the population of Black Americans increases (decreases), the level of energy expenditure increases (decreases).*

Let's discuss what we've found as a class:

Questions for the Class

- What can we say about our hypothesis? Does it depend on U.S. State?
- How would you answer our research question now that we have analyzed the data?
- What can we conclude about the relationship between sustainability and disadvantaged communities? What might you recommend from a policy-making perspective?
- **Can** we make sound policy decisions, based on what we've seen so far?

NOTE: Please do not go on to the next page until your instructor tells you to!

Differences between Frequentist & Bayesian Approaches

There are key differences between Frequentist and Bayesian approaches. In the activity, we focused on model summaries and general inference.

General Inference

For general inference, we are concerned with answering questions about a *larger* dataset, while only having access to a *smaller* dataset.

1. Translate our question into a mathematical model, which requires assumptions
2. Fit the model using data
3. Interpret model results in terms of our question

Table 1: General Inference Comparison

Frequentist	Bayesian
Provides: Pr(data | hypothesis)	Provides: Pr(hypothesis | data)
Decision based on hard cut-off	Decision based on posterior distribution

Model Summaries

- Statistical models are complex objects, so we use *model summaries* to help.
- Statistical models represent the *uncertainty due to limited data.*

Some common summaries:

Table 2: Common Model Summaries

Point estimate	Uncertainty	Interval
42	4	(42 +/- 2x4) = (34, 50)

Here are the differences between how Frequentists and Bayesians get model summaries:

Table 3: Model Summary Comparison

Frequentist	Bayesian
Maximum likelihood estimate	Maximum of the posterior density
Standard error	Posterior standard deviation
Confidence interval	Credible interval

Applying the Differences

With the remainder of class time, separate into groups of 2-3 people, where each group should have at least 1 person who did the Frequentist analysis and 1 person who did the Bayesian analysis. Using the critical differences listed on the previous page, discuss the following:

- How do our modeling choices affect our outcomes of interest based on the assumptions we know about both approaches? Are the results for our hypothesis different? If so, how?
- How do we interpret the coefficients (m and b) in both the Frequentist and Bayesian models?
- How might we improve our model based on what we know from our **model summaries** and **general inference** for both approaches?

Jot down 2-3 sentences for each question, as well as any remaining questions or concerns you have about conducting your analyses using one of the two approaches.

NOTE: Please do not flip to the last page until your instructor tells you to.

Please use this QR code to access the post-activity survey.