Course Project
Exploratory Data Analysis for Machine Learning
Ivan Patel
12/19/2020

Introduction

In this course project, I wanted to understand the American public's support of energy production from different sources. I utilize the Pew Research survey on energy resources (2018) dataset downloaded from the website of OpenIntro.[1] After downloading the dataset, I ran hypothesis testing on the proportion of respondents who favored expanding offshore drilling.

Per the University of Calgary, offshore drilling is a process of extracting petroleum from reserves located beneath oceans.[2] It is one of the energy industry techniques to meet the growing demand for liquid fuels. Since 2015 Q1, both world production and consumption of liquid fuels has steadily increased. [3] Despite the sharp decline in 2020, the administration projects that in 2021 the total world consumption of petroleum and other liquids will increase to 98.16 million barrels per day.

Despite the ability of offshore drilling to meet world demands for oil, it possesses severe risks to marine life.[4] Moreover, incidents such as the Deepwater Horizon explosion or oil spills could have made the American public against offshore drilling. For this project, I wanted to know if most Americans favored the expansion of offshore drilling?

Brief Description of the data set and a summary of its attributes

The final dataset consists of 2541 observations and six columns for six different energy sources: solar panel farms, wind turbine farms, offshore drilling, hydraulic fracturing, coal mining, and nuclear power plants. The dataset had no missing observations.

To obtain this data, the Pew Research Center used a nationally representative panel of randomly selected U.S. adults living in a household recruited from landline and cellphone random-digit-dial surveys. Participants that did not have an internet connection were provided a tablet and wireless connection.

Summary Statistics

	solar_panel_farms	wind_turbine_farms	offshore_drilling	hydrolic_fracturing	coal_mining	nuclear_power_plants
count	2541	2541	2541	2541	2541	2541
unique	2	2	2	2	2	2
top	favor_increase	favor_increase	no_increase	no_increase	no_increase	no_increase
freq	2261	2160	1550	1550	1601	1423

[1] https://www.openintro.org/data/index.php?data=pew_energy_2018
[2] https://energyeducation.ca/encyclopedia/Offshore_drilling
[3] https://www.eia.gov/outlooks/steo/report/global_oil.php
[4] https://usa.oceana.org/our-campaigns/offshore_drilling/campaign

Key Findings and Insights

When exploring the data, the first thing I noticed was that all columns are categorical. Each column for any observation could take on one of the two possible values: favor increase or no increase. The offshore drilling bar plot in the Appendix shows that about 39% of respondents favored the increase in offshore drilling usage. This sample proportion is our estimate of the actual percentage of Americans who favored the rise in the use of offshore drilling. The other bar plots show that solar panel farms and opposing coal mining increase had overwhelming support.

However, if we reproduced this study by following the same methodology (random sampling), we would have contacted a different group of people. Had another sample been drawn, the sample proportion in favor of offshore drilling would have been different. Thus, the ratio differs from one random sample to the next. The distribution of these sample proportions is a sampling distribution, and we will use it to do our hypothesis testing. [5]

When our sample is randomly selected, and the sample size is sufficiently large, our sampling distribution of the sample proportion, $\hat{p}$, will follow a normal distribution with the following mean and standard error:

$\mu_{\hat{p}} = p \qquad SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion. We can use our hypothesized value to compute the sampling distribution's standard error when the null hypothesis is correct.[6]

Formulating at least three hypotheses about this data

Given the data, one can use hypothesis testing to answer any or all the following three questions:

1. Do the majority of Americans favor the increase in offshore drilling?

2. Do the majority of Americans oppose or favor the expansion of coal mining to produce energy?

3. Do the majority of Americans favor the increase in the usage of solar panel farms to produce energy?

Significance test for one of three hypotheses and discussing the results.

I test the first hypothesis. If p is the proportion of Americans that favor the increase in offshore drilling, then

$$H_0: p = 0.5$$
$$H_A: p < 0.5$$

Per Research's sample shows that 39% of American adults favor the increase in offshore drilling. Does this 39% represent a real difference from the null hypothesis of 50%? In other words, is our

[5] Stock, J. H., and M. W. Watson. 2015. *Introduction to Econometrics, Third Update, Global Edition*. Pearson Education Limited.

[6] https://leanpub.com/openintro-statistics

sample proportion of 39% statistically significant at a 5% significance level under the null hypothesis?

As mentioned before, if the null hypothesis is correct, then the sampling distribution indicates that a sampling proportion based on a large sample size would be normally distributed. Our sampling distribution's standard error under the null hypothesis is $\sqrt{\frac{0.5(1-0.5)}{2541}} \approx 0.0091.$

Using this, we can compute the Z-score of our sample proportion, 0.39. This Z-score tells us how many standard deviations away from the mean is sample proportion on 0.39.

$$Z = \frac{0.39 - 0.5}{0.0091} \approx -11.09.$$

We can use a Z table to find the p-value of our Z-score. Although -11.09 is off the table, we could use the smallest area listed: 0.0002. Notice that the p-value is less than our significance level of 5%. [7]

If the null hypothesis were correct, there is a minimal chance of observing such an extreme deviation from 0.5. Thus, we can reject the null hypothesis, and our data provide enough evidence that less than a majority of Americans favor the increase in offshore drilling.

Suggestions for next steps in analyzing this data

The next steps in analyzing this data would be to formulate additional hypotheses about the other columns and conduct a formal significance test for those statements. For example, one can answer, do the majority of Americans support solar panels, or is the support greater than 0.5?

Summarizing the quality of this data set

Although this data set is impressive, it does have its limitations. For example, this dataset is more than two years old, and recent survey results should be used to understand if the opinions expressed by this sample hold for the current population.

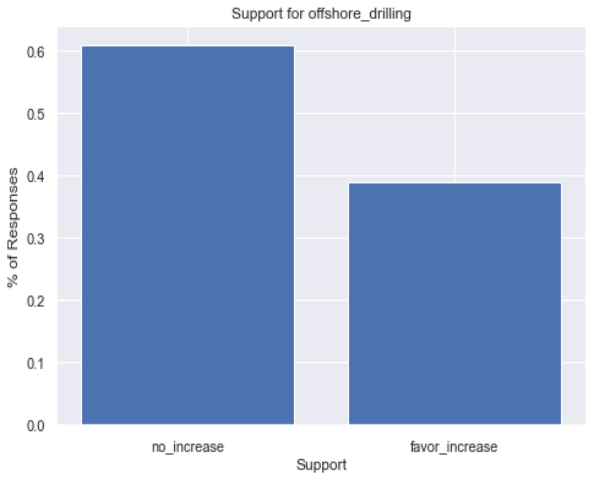
Moreover, "the authors did not have access to individual responses in the original data set. Instead, they took the published percentages and backed out the breakdown." Not having individual responses from the original dataset collected and analyzed by the Pew Research Center could bias our results.

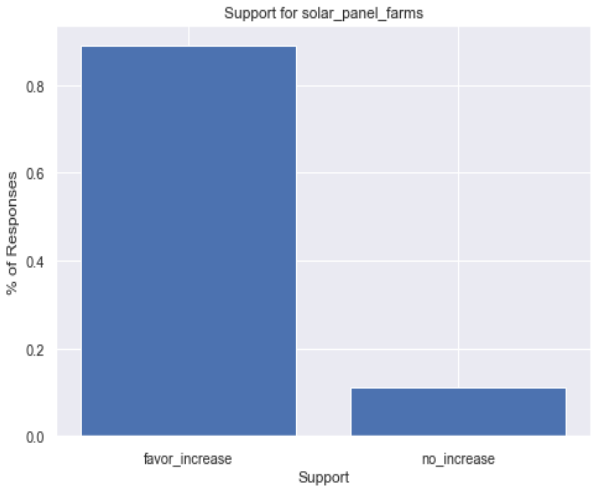
We also do not know if the questions' framing evoked biased or specific responses that neutral questions would not have brought about.

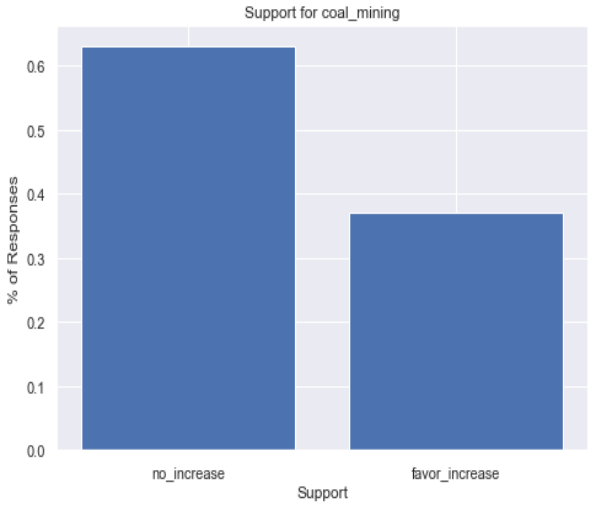
Finally, responses to all energy sources fell only in two categories. I wish there were one more category such as 'favor_decrease' because some Americans might want to shrink the energy production from specific sources.

[7] http://www.z-table.com/

Appendix


Support for offshore_drilling
% of Responses
0.0
0.1
0.2
0.3
0.4
0.5
0.6
no_increase
favor_increase
Support


Support for solar_panel_farms
% of Responses
0.0
0.2
0.4
0.6
0.8
favor_increase
no_increase
Support


Support for coal_mining
% of Responses
0.0
0.1
0.2
0.3
0.4
0.5
0.6
no_increase
favor_increase
Support