**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]](#footnote-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]](#footnote-2) It is one of the processes used by the energy industry to meet the growing demand of liquid fuels. Since 2015 Q1, both world production and consumption of liquid fuels has steadily increased. [[3]](#footnote-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 advantages of offshore drilling, it possesses severe risks to marine life.[[4]](#footnote-4) Moreover, incidents such as Deepwater Horizon explosion or oil spills could have affected public opinion on offshore drilling. For this project, my research question is do majority of Americans favor 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 6 columns for six different sources of energy: 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 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**

**Key Findings and Insights**

When exploring the data, the first thing I noticed that was 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 usage of offshore drilling. This sample proportion is our estimate of the actual percentage of Americans who favored the increase in usage of offshore drilling. The other bar plots show that for solar panel farms and opposing coal mining increase had an 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 proportion 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]](#footnote-5)

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

, where p is the population proportion. We can use our hypothesized value to compute the standard error of the sampling distribution when the null hypothesis is true.[[6]](#footnote-6)

**Formulating at least 3 hypotheses about this data**

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

1. Do majority of Americans favor the increase in offshore drilling?
2. Do majority of Americans oppose or favor the expansion of coal mining to produce energy?
3. Do majority of American favor the increase the usage of solar panel farms to produce energy?

**Significance test for one of three hypothesis and discussing the results.**

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

Per Research’s sample shows that 39% of American adults favor the increase in offshore drilling. Does this 39% represent a real difference from null hypothesis of 50%? In other words, is our sample proportion of 39% statistically significant at a 5% significance level under the null hypothesis?

As mentioned before, if the null hypothesis is true, then the sampling distribution indicates that a sampling proportion based on large sample size would be normally distributed. The standard error of our sampling distribution under the null hypothesis is

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.

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]](#footnote-7)

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

**Suggestions for next steps in analyzing this data**

The next steps in analyzing this data further would be to formulate additional hypotheses about the other columns and conducing a formal significance test for those statements. For example, one can answer do 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 interesting, it does have its limitations. For example, this dataset is more than 2 years old and a 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 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 framing of the questions evoked biased responses that neutral questions would not have brought about.

Finally, responses on all energy sources fell only in two categories. I wish there was one more category such as ‘favor\_decrease’ because some Americans might want to shrink the energy production from certain sources.

**Appendix**

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

1. https://www.openintro.org/data/index.php?data=pew\_energy\_2018 [↑](#footnote-ref-1)
2. https://energyeducation.ca/encyclopedia/Offshore\_drilling [↑](#footnote-ref-2)
3. https://www.eia.gov/outlooks/steo/report/global\_oil.php [↑](#footnote-ref-3)
4. https://usa.oceana.org/our-campaigns/offshore\_drilling/campaign [↑](#footnote-ref-4)
5. Stock, J. H., and M. W. Watson. 2015. *Introduction to Econometrics, Third Update, Global Edition*. Pearson Education Limited. [↑](#footnote-ref-5)
6. https://leanpub.com/openintro-statistics [↑](#footnote-ref-6)
7. http://www.z-table.com/ [↑](#footnote-ref-7)