project2

April 11, 2020

1 Project 2: Diet and Disease

Your first and last name: Naima Amraan

Project parter's first and last name: ...

Partners. Undergraduate students may work with one other partner. Both of you are required to submit the project. Please designate your partner's name so we know with whom you worked.

In this project, you will investigate the major causes of death in the world, as well as how one of these causes, heart disease, might be linked to diet!

1.0.1 Logistics

Deadline. This project is due at 11:59pm on Friday, 4/10. It's **much** better to be early than late, so start working now.

Checkpoint. Complete the questions up until the end of Part 2 and submit them by 11:59pm on Friday, 4/3. This will carry the weight of one homework assignment.

Partners. Undergraduate students may work with a partner. Only one of you is required to submit the project. The person who submits should also designate their partner so that both receive credit.

Rules. Don't share your code with anybody but your partner. You are welcome to discuss questions with other students, but don't share the answers. The experience of solving the problems in this project will prepare you for exams (and life). If someone asks you for the answer, resist! Instead, you can demonstrate how you would solve a similar problem.

Support. You are not alone! Come to office hours, post on Piazza, and talk to your classmates. If you want to ask about the details of your solution to a problem, make a private Piazza post and the staff will respond. If you're ever feeling overwhelmed or don't know how to make progress, be sure to come to office hours and speak to a TF, ULA, or instructor.

Advice. Develop your answers incrementally. To perform a complicated table manipulation, break it up into steps, perform each step on a different line, give a new name to each result, and check that each intermediate result is what you expect. You can add any additional names or functions you want to the provided cells.

All of the concepts necessary for this project are found in the textbook. If you are stuck on a particular problem, reading through the relevant textbook section often will help clarify the concept.

To get started, load `datascience`, `numpy`, and `plots`.

Credit: This project has been adapted from Berkeley's Data8 course.

[93]:
```
from datascience import *
import numpy as np

%matplotlib inline
import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
np.set_printoptions(legacy='1.13')
```

2 Diet and Cardiovascular Disease

Death and its many causes are often a disconcerting topic for polite conversation. However, the more we know about it, the better equipped we are to prevent our early demise. As the acclaimed Professor Albus Dumbledore once said, "After all, to the well-organized mind, death is but the next great adventure."

In the following analysis, we will investigate the world's most dangerous killer: Cardiovascular Disease. Your investigation will take you across decades of medical research, and you'll look at multiple causes and effects across two different studies.

Here is a roadmap for this project:

- In Part 1, we'll investigate the major causes of death in the world during the past century (from 1900 to 2015).
- In Part 2, we'll look at data from the Framingham Heart Study, an observational study of cardiovascular health.
- In Part 3, we'll examine the clinical trials from the Minnesota Coronary Experiment and introduce our second dataset.
- In Part 4, we'll run a hypothesis test on our observed data from the Minnesota Coronary Experiment.
- In Part 5, we'll conclude the experiment and reflect on what we've learned about the relationship between diet and cardiovascular disease.

2.1 Part 1: Causes of Death

In order to get a better idea of how we can most effectively prevent deaths, we need to first figure out what the major causes of death are. Run the following cell to read in and view the `causes_of_death` table, which documents the death rate for major causes of deaths over the last century (1900 until 2015).

[94]:
```
causes_of_death = Table.read_table('causes_of_death.csv')
causes_of_death.show(5)
```

```
<IPython.core.display.HTML object>
```

Each entry in the column **Age Adjusted Death Rate** is a death rate for a specific **Year** and **Cause** of death.

The **Age Adjusted** specification in the death rate column tells us that the values shown are the death rates that would have existed if the population under study in a specific year had the same

age distribution as the "standard" population, a baseline. This is so we can compare ages across years without worrying about changes in the demographics of our population.

Question 1.1. What are all the different causes of death in this dataset? Assign an array of all the unique causes of death to `all_unique_causes`.

[95]:
```
all_unique_causes = causes_of_death.group ('Cause').column ('Cause')
sorted(all_unique_causes)
```

[95]:
```
['Accidents', 'Cancer', 'Heart Disease', 'Influenza and Pneumonia', 'Stroke']
```

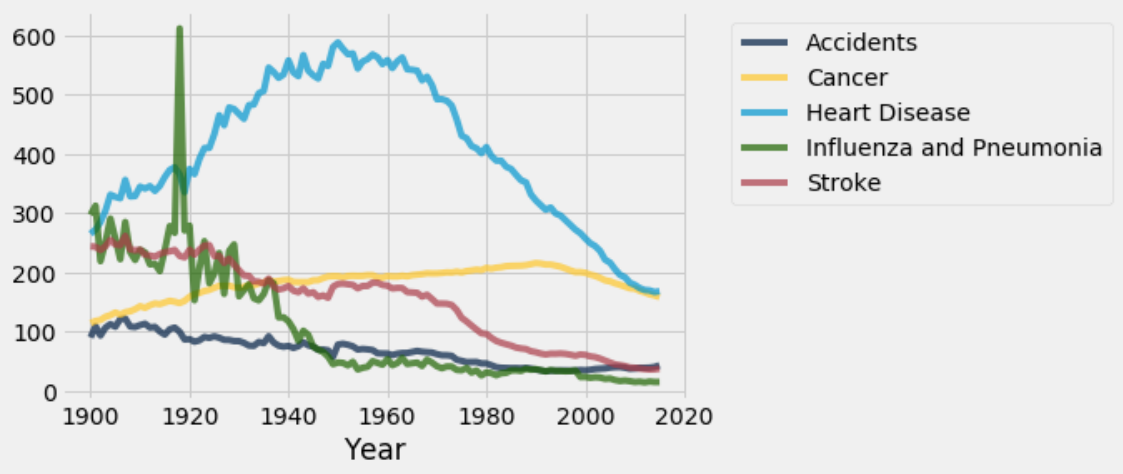
Question 1.2. We would like to plot the death rate for each disease over time. To do so, we must create a table with one column for each cause and one row for each year.

Create a table called `causes_for_plotting`. It should have one column called `Year`, and then a column with age-adjusted death rates for each of the causes you found in Question 1.1. There should be as many of these columns in `causes_for_plotting` as there are causes in Question 1.1.

Hint: Use `pivot`, and think about how the `elem` function might be useful in getting the **Age Adjusted Death Rate** for each cause and year combination.

[96]:
```
def elem(x):
    return x.item(0)
```

[97]:
```
causes_for_plotting = causes_of_death.pivot('Cause', 'Year', values='Age␣
↪Adjusted Death Rate', collect=elem)

# Do not change this line
causes_for_plotting.plot('Year')
```



Let's examine the graph above. You'll see that in the 1960s, the death rate due to heart disease steadily declines. Up until then, the effects of smoking, blood pressure, and diet on the cardiovascular system were unknown to researchers. Once these factors started to be noticed, doctors were able recommend a lifestyle change for at-risk patients to prevent heart attacks and heart problems.

Note, however, that the death rate for heart disease is still higher than the death rates of all other causes. Even though the death rate is starkly decreasing, there's still a lot we don't understand about the causes (both direct and indirect) of heart disease.

2.2 Part 2: The Framingham Heart Study

The Framingham Heart Study is an observational study of cardiovascular health. The initial study followed over 5,000 volunteers for several decades, and followup studies even looked at their descendants. In this section, we'll investigate some of its key findings about diet, cholesterol, and heart disease.

Run the cell below to examine data for almost 4,000 subjects from the first wave of the study, collected in 1956.

[98]:
```
framingham = Table.read_table('framingham.csv')
framingham
```

[98]:
```
AGE  | SYSBP | DIABP | TOTCHOL | CURSMOKE | DIABETES | GLUCOSE | DEATH | ANYCHD
39   | 106   | 70    | 195     | 0        | 0        | 77      | 0     | 1
46   | 121   | 81    | 250     | 0        | 0        | 76      | 0     | 0
48   | 127.5 | 80    | 245     | 1        | 0        | 70      | 0     | 0
61   | 150   | 95    | 225     | 1        | 0        | 103     | 1     | 0
46   | 130   | 84    | 285     | 1        | 0        | 85      | 0     | 0
43   | 180   | 110   | 228     | 0        | 0        | 99      | 0     | 1
63   | 138   | 71    | 205     | 0        | 0        | 85      | 0     | 1
45   | 100   | 71    | 313     | 1        | 0        | 78      | 0     | 0
52   | 141.5 | 89    | 260     | 0        | 0        | 79      | 0     | 0
43   | 162   | 107   | 225     | 1        | 0        | 88      | 0     | 0
... (3832 rows omitted)
```

Each row contains data from one subject. The first seven columns describe the subject at the time of their initial medical exam at the start of the study. The last column, `ANYCHD`, tells us whether the subject developed some form of heart disease at any point after the start of the study.

You may have noticed that the table contains fewer rows than subjects in the original study: this is because we are excluding subjects who already had heart disease as well as subjects with missing data.

2.2.1 Diabetes and the population

Before we begin our investigation into cholesterol, we'll first look at some limitations of this dataset. In particular, we will investigate ways in which this is or isn't a representative sample of the population by examining the number of subjects with diabetes.

According to the CDC, the prevalence of diagnosed diabetes (i.e., the percentage of the population who have it) in the U.S. around this time was 0.93%. We are going to conduct a hypothesis test with the following null and alternative hypotheses:

Null Hypothesis: The probability that a participant within the Framingham Study has diabetes is equivalent to the prevalence of diagnosed diabetes within the population. (i.e., any difference is due to chance).

Alternative Hypothesis: The probability that a participant within the Framingham Study has diabetes is different than the prevalence of diagnosed diabetes within the population.

We are going to use the absolute distance between the observed prevalence and the true population prevalence as our test statistic. The column `DIABETES` in the `framingham` table contains a 1 for subjects with diabetes and a 0 for those without.

Question 2.1. What is the observed value of the statistic in the data from the Framingham Study? You should convert prevalences to proportions before calculating the statistic!

[99]:
```
observed_prevalence = sum(framingham.column (5))/framingham.num_rows
observed_diabetes_distance = abs(observed_prevalence-0.0093)
observed_diabetes_distance
```

[99]:
```
0.018029515877147319
```

[ ]:
```

```

Question 2.2. The array `diabetes_proportions` contains the proportions of the population without and with diabetes. Complete the following code to simulate 5000 values of the statistic under the null hypothesis.

[100]:
```
diabetes_proportions = make_array(.9907, .0093)

diabetes_simulated_stats = make_array()
repetitions = 5000

for i in np.arange(repetitions):
    simulated_stat = sample_proportions(4000, diabetes_proportions).item(1)
    diabetes_simulated_stats = np.append (diabetes_simulated_stats,␣
 ↪simulated_stat)

diabetes_simulated_stats
```

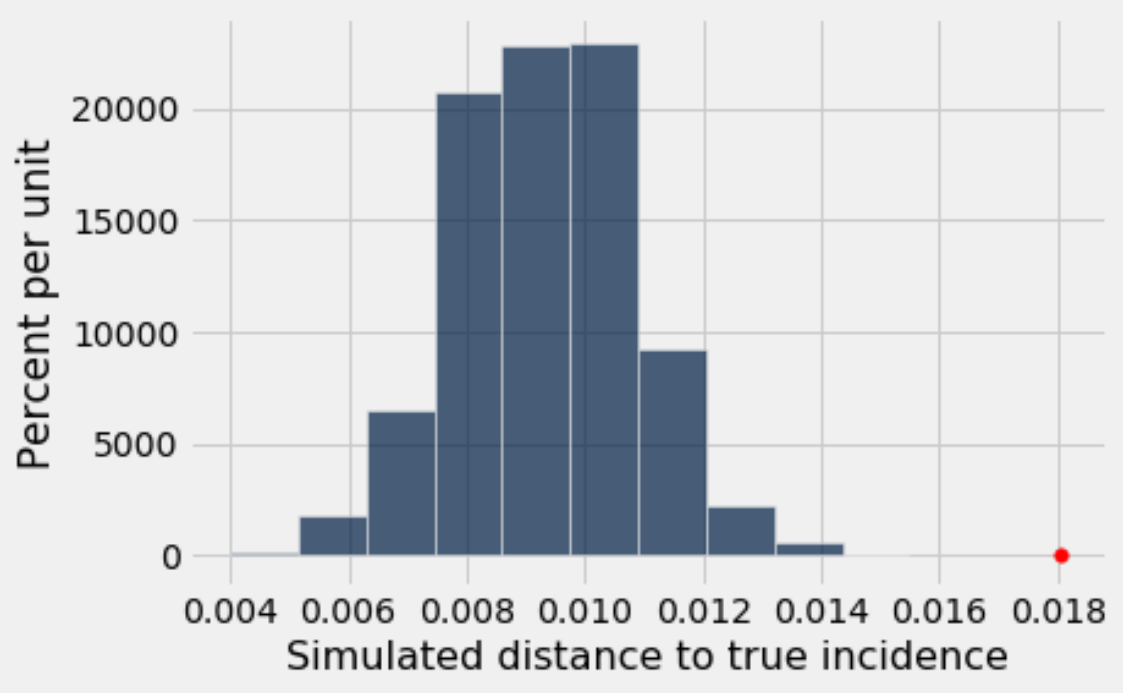
[100]:
```
array([ 0.00975,  0.01125,  0.009  , ...,  0.011  ,  0.0075 ,  0.0095 ])
```

Question 2.3. Run the following cell to generate a histogram of the simulated values of your statistic, along with the observed value.

Make sure to run the cell that draws the histogram, since it will be graded.

[101]:
```
Table().with_column('Simulated distance to true incidence',␣
 ↪diabetes_simulated_stats).hist()
plots.scatter(observed_diabetes_distance, 0, color='red', s=30)
```

[101]:
```
<matplotlib.collections.PathCollection at 0x7f52b3af6550>
```



Question 2.4. Based on the results of the test and the empirical distribution of the test statistic under the null, should you reject the null hypothesis?

Yes you should reject the null hypothesis

Question 2.5. Suppose you know that the study was well-designed to represent the population. Why might there be a difference between the population and the sample?

Perhaps the subjects under the study were more likely than the average people to develop diabetes i.e. had a high sugar intake in their meals than the average people or perhaps they were older people who are more likely to develop diabetes

In real-world studies, getting a truly representative random sample of the population is often incredibly difficult. Even just to accurately represent all Americans, a truly random sample would need to examine people across geographical, socioeconomic, community, and class lines (just to name a few). For a study like this, scientists would also need to make sure the medical exams were standardized and consistent across the different people being examined. In other words, there's a tradeoff between taking a more representative random sample and the cost of collecting all the data from the sample.

The Framingham study collected high-quality medical data from its subjects, even if the subjects may not be a perfect representation of the population of all Americans. This is a common issue that data scientists face: while the available data aren't perfect, they're the best we have. The Framingham study is generally considered the best in its class, so we'll continue working with it while keeping its limitations in mind.

(For more on representation in medical study samples, you can read these recent articles from NPR and Scientific American).

2.2.2 Section 2: Cholesterol and Heart Disease

In the remainder of this part, we are going to examine one of the main findings of the Framingham study: an association between serum cholesterol (i.e., how much cholesterol is in someone's blood) and whether or not that person develops heart disease.

We'll use the following null and alternative hypotheses:

Null Hypothesis: In the population, the distribution of cholesterol levels among those who get heart disease is the same as the distribution of cholesterol levels among those who do not.

Alternative Hypothesis: The cholesterol levels of people in the population who get heart disease are higher, on average, than the cholesterol level of people who do not.

Question 2.6. From the provided Null and Alternative Hypotheses, what seems more reasonable to use, A/B Testing or the Standard Hypothesis Testing? Assign the variable `reasonable_test` to one of the following choices.

1. A/B Testing
2. Standard Hypothesis Test

[102]:
```
reasonable_test = 1
reasonable_test
```

[102]:
```
1
```

Question 2.7. Now that we have a null hypothesis, we need a test statistic. Explain and justify your choice of test statistic in two sentences or less.

Hint: Remember that larger values of the test statistic should favor the alternative over the null.

**Test Statistic: Absolute difference of the means between the cholesterol levels of people with and without a heart disease since the column 'ANYCHD' does not have an equal number of 0s(without) and 1s (with)

Question 2.8. Write a function that computes your test statistic. It should take a table with two columns, `TOTCHOL` and `ANYCHD`, and compute the test statistic you described above.

[103]:
```
def compute_framingham_test_statistic(framingham):
    ch_with_hd = framingham. where ('ANYCHD', 1).column ('TOTCHOL')
    ch_without_hd = framingham. where ('ANYCHD', 0).column ('TOTCHOL')
    test_stat = np.average (ch_with_hd) - np.average(ch_without_hd)
    return test_stat
```

Question 2.9. Use the function you defined above to compute the observed test statistic, and assign it to the name `framingham_observed_statistic`.

[104]:
```
framingham_observed_statistic = compute_framingham_test_statistic(framingham)
framingham_observed_statistic
```

[104]:
```
16.635919905689406
```

Now that we have defined hypotheses and a test statistic, we are ready to conduct a hypothesis test. We'll start by defining a function to simulate the test statistic under the null hypothesis, and then use that function 1000 times to understand the distribution under the null hypothesis.

Question 2.10. Write a function to simulate the test statistic under the null hypothesis.

The `simulate_framingham_null` function should simulate the null hypothesis once (not 1000 times) and return the value of the test statistic for that simulated sample.

[105]:
```
def simulate_framingham_null():
    shuffled_frame = framingham. sample (with_replacement = False).column␣
 ↪('TOTCHOL')
    sim_table_frame = framingham. with_column('TOTCHOL', shuffled_frame)
    return compute_framingham_test_statistic(sim_table_frame)
```

[106]:
```
# Run your function once to make sure that it works.
simulate_framingham_null()
```

[106]:
```
1.6731664895439451
```

Question 2.11. Fill in the blanks below to complete the simulation for the hypothesis test. Your simulation should compute 1000 values of the test statistic under the null hypothesis and store the result in the array framingham_simulated_stats.

Hint: You should use the function you wrote above in Question 2.10.

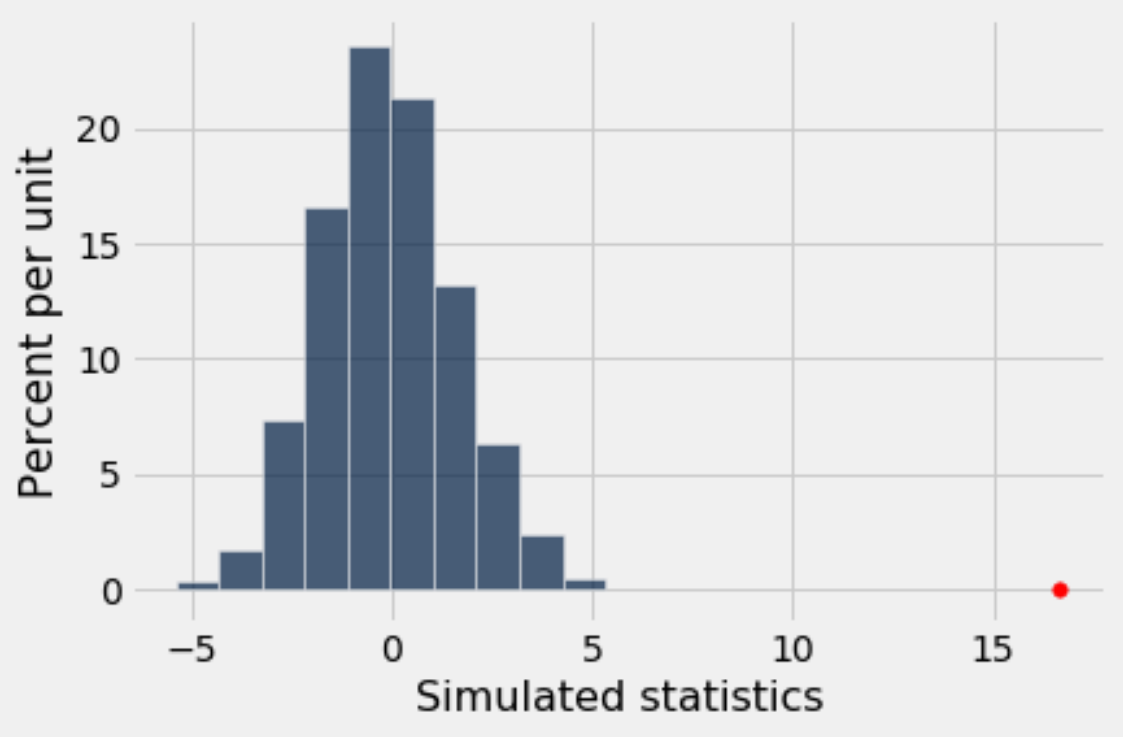
Note: Warning: running your code might take a few minutes! We encourage you to check your `simulate_framingham_null()` code to make sure it works correctly before running this cell.

[107]:
```
framingham_simulated_stats = make_array()

for i in range (1000):
    sim_stat = simulate_framingham_null()
    framingham_simulated_stats = np.append (framingham_simulated_stats,␣
 ↪sim_stat)
```

The following line will plot the histogram of the simulated test statistics, as well as a point for the observed test statistic. Make sure to run it, as it will be graded.

[108]:
```
Table().with_column('Simulated statistics', framingham_simulated_stats).hist()
plots.scatter(framingham_observed_statistic, 0, color='red', s=30)
```

[108]:
```
<matplotlib.collections.PathCollection at 0x7f52b3a0ecc0>
```



Question 2.12. Compute the p-value for this hypothesis test, and assign it to the name `framingham_p_value`.

Hint: One of the key findings of the Framingham study was a strong association between cholesterol levels and heart disease. If your p-value doesn't match up with this finding, you may want to take another look at your test statistic and/or your simulation.

[109]:
```
framingham_p_value = np.count_nonzero(framingham_simulated_stats >=␣
↪framingham_observed_statistic)
framingham_p_value
```

[109]: 0

Question 2.13. Despite the Framingham Heart Study's reputation as a well-conducted and rigorous study, it has some major limitations. Give one specific reason why it can't be used to say that high cholesterol *causes* heart disease.

The Framigham study was an observational study where they observed an association between high cholesterol and heart disease. This is an issue of causality; the fact that high cholesterol is associated with heart disease it does mean that it causes heart disease.

Similar studies from the 1950s found positive associations between diets high in saturated fat, high cholesterol, and incidence of heart disease. In 1962, the U.S. Surgeon General said:

"Although there is evidence that diet and dietary habits may be implicated in the development of coronary heart disease and may be significant in its prevention or control, at present our only research evidence is associative and not conclusive."

2.3 Part 3: Causality, the National Diet-Heart Study, and the Minnesota Coronary Experiment

To establish a causal link between saturated fat intake, serum cholesterol, and heart disease, a group of doctors in the US established the National Heart-Diet Study. The study was based in 6 centers: Baltimore, Boston, Chicago, Minneapolis-St. Paul, Oakland, and Faribault, MN. The first 5 centers recruited volunteers from the local population: volunteers and their families were asked to adjust their diet to include more or less saturated fat.

You may already have a strong intuition about what the doctors concluded in their findings, but the evidence from the trial was surprisingly complex.

Question 3.1. Why might the data from the National Heart-Diet Study not be enough to determine causality? Describe one specific limitation of the data from these first 5 centers in the study.

Hint: what is the main problem with fad diets?

Because fad diets work for a short amount of time hence is not enough to determine causality of something as long term as heart disease.

The sixth center was organized by Dr. Ivan Frantz, and its study was known as the Minnesota Coronary Experiment. Dr. Frantz was a strong proponent of reducing saturated fats to prevent death from heart disease. He believed so strongly in the idea that he placed his household on a strict diet very low in saturated fats. The main difference between the Minnesota Coronary Experiment and the rest of the National Diet-Heart Study was the setting. While the other centers in the study looked at volunteers, Dr. Frantz conducted his study at Faribault State Hospital, which housed patients who were institutionalized due to disabilities or mental illness.

In this institution, the subjects were randomly divided into two equal groups: half of the subjects, the **control group**, were fed meals cooked with saturated fats, and the other half, the **diet group**, were fed meals cooked with polyunsaturated fats. For example, the diet group's oils were replaced with corn oils and their butter was replaced with margarine. The subjects did not know which food they were getting, to avoid any potential bias or placebo effect. This type of study is known as a **blind** study.

Although standards for informed consent in participation weren't as strict then as they are today, the study was described as follows:

"No consent forms were required because the study diets were considered to be acceptable as house diets and the testing was considered to contribute to better patient care. Prior to beginning the diet phase, the project was explained and sample foods were served. Residents were given the opportunity to decline participation."

Despite the level of detail and effort in the study, the results of the study were never extensively examined until the late 20th century. Over 40 years after the data were collected, Dr. Christopher Ramsden heard about the experiment, and asked Dr. Frantz's son Robert to uncover the files in the Frantz family home's dusty basement. You can learn more about the story of how the data was recovered on the Revisionist History podcast or in Scientific American magazine.

Question 3.2. While the data from such a study may be useful scientifically, it also raises major ethical concerns. Describe at least one ethical problem with the study conducted at Faribault State Hospital.

Hint: There isn't necessarily a single right or wrong answer to this question. If you're not sure, some areas of consideration may be the study organizers' selection of participants for the study, as well as their justification for not using consent forms. You could also ask yourself how the project might have been explained to the patients prior to the diet phase, and to what degree were they capable of consent.

Patients with mental disabilities may not be as capable to understand information provided to them as part of the research hence affecting consent that they give.

In recent years, poor treatment of patients at Faribault State Hospital (and other similar institutions in Minnesota) has come to light: the state has recently changed patients' gravestones from numbers to their actual names, and apologized for inhumane treatment of patients.

Unfortunately, the data for each individual in the 1968 study is not available; only summary statistics are available. Therefore, in this project we create artificial synthetic data, based on those summary statistics.

In order to test whether following the diet actually reduced serum cholesterol levels, we need to create a table with one row for each participant in the study, as well as how their serum cholesterol changed. There were 1179 subjects in the diet group and 1176 subjects in the control group who had their serum cholesterol changes measured.

The study measured the serum cholesterol at the start and end of the study, then used this to compute the percentage change for each individual. Then, they computed the average and standard deviation of these percentage changes for each study group. We have these summary statistics: for those who received the unsaturated fat diet, the serum cholestorol decreased by 13.8% on average, with a standard deviation of 13%. For those in the control group, the percentage change decreased by 1% on average, with a standard deviation of 14.5%. We used these statistics to generate random synthetic percentage change levels for each individual, making an assumption about the distribution fo these changes. We have saved this data in `serum_cholesterol.csv`. We read this table into `serum_cholesterol` below.

[110]:
```
serum_cholesterol = Table.read_table('serum_cholesterol.csv')
serum_cholesterol
```

[110]:
```
Condition | Change in Serum Cholesterol
Diet      | -8.36662
Diet      | -23.6885
Diet      | -28.985
Diet      | -10.9341
Diet      | -17.9041
Diet      | -11.7145
Diet      | -13.6215
Diet      | -2.2387
Diet      | -2.03579
Diet      | -13.5746
... (2345 rows omitted)
```

After determining if serum cholesterol is actually lowered by this new diet, we will see whether or not death rates were reduced as well. The following table is a summarized version of the data collected in the experiment.

[111]:
```
mortality_summary = Table.read_table('mortality_summary.csv')
mortality_summary
```

[111]:
```
Age   | Condition | Total | Deaths | CHD Deaths
0-34  | Diet      | 1367  | 3      | 0
35-44 | Diet      | 728   | 3      | 0
45-54 | Diet      | 767   | 14     | 4
55-64 | Diet      | 870   | 35     | 7
```

```
65+   | Diet      | 953   | 190    | 42
0-34  | Control   | 1337  | 7      | 1
35-44 | Control   | 731   | 4      | 1
45-54 | Control   | 816   | 16     | 4
55-64 | Control   | 896   | 33     | 12
65+   | Control   | 958   | 162    | 34
```

Question 3.3. The numbers of deaths in the Deaths column above are not specific to cardiovascular disease. For our tests, we are going to use the total number of deaths instead of the number of CHD deaths. If a hypothesis test shows that the rate of deaths in the diet group is different from the rate of deaths in the control group, which of the following are valid conclusions from the test? Assign the name `mortality_valid_conclusions` to a list of numbers.

1. Eating a diet rich in unsaturated fats causes an increased/decreased risk of death.
2. Eating a diet rich in unsaturated fats causes/prevents cardiovascular disease.
3. Lower cholesterol causes an increased/decreased risk of cardiovascular disease.
4. It is impossible to determine any causal relationship between any of these factors, even if the test shows an association.

[112]:
```
mortality_valid_conclusions = [1]
mortality_valid_conclusions
```

[112]:
```
[1]
```

To help with our simulations, we are going to expand the `mortality_summary` table so that we have one row for every subject in the experiment. Our goal is to put this into a table called `minnesota_data`.

Question 3.4. Using all of the notes below, complete the code below to create a table with four columns: "Age", "Condition", "Participated" and "Died". Each row should contain a specific patient and should have their age group and condition as specified in the `mortality_summary` table, a `True` in the "Participated" column since everyone participated in the experiment, and either a `True` or `False` in the "Died" column, depending on if they are alive or dead.

The total number of rows of `minnesota_data` should be the same as the number of participants summarized in the mortality_summary table.

Hint: The most useful notes from below will be the final three; how to get an item out of a row, passing in just one value into the second argument of `with_column`, and how to iterate over rows. Make sure you use the other two notes to understand what the rest of this code is doing.

The following few notes will all be helpful to finish and understand the code below:

- `tbl1.append(tbl2)` adds all of the rows of `tbl2` into `tbl1`, assuming they have the same column names
- `np.arange(5) < 3` returns the following array: `[True, True, True, False, False]`
- `row.item(x)` returns the item in column `x` in a specific row of a table
- If `my_table` has 10 rows. Then, `my_table.with_column('Num', val)` adds an array of length 10, with each element being val, as a new column of the table.
- To iterate over all rows of a table, you can write `for row in tbl.rows:`

[114]:
```
minnesota_data = Table(['Age', 'Condition', 'Died', 'Participated'])

for row in  mortality_summary.rows:
```

```
    i = np.arange(0, row.item('Total'))
    t = Table().with_column('Died', i < row.item('Deaths'))
    t = t.with_column('Age', np.array([row.item('Age')]*row.item('Total')) )
    t = t.with_column('Condition', np.array([row.item('Condition')]*row.
 ↪item('Total')))
    t = t.with_column('Participated', True)
    minnesota_data.append(t)

minnesota_data
```

[114]:
```
Age   | Condition | Died   | Participated
0-34 | Diet      | True   | True
0-34 | Diet      | True   | True
0-34 | Diet      | True   | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
... (9413 rows omitted)
```

2.4 Part 4: Running a Hypothesis Test

Now that we have two clean datasets from the Minnesota Coronary Experiment to work with, we can focus on determining causal links. Assuming that these randomized controlled experiments are samples from the larger population, we can work on using the inference techniques discussed so far in the course to answer the following questions:

- Does changing saturated fats to polyunsaturated fats in a person's diet **decrease their serum cholestrol levels**?
- Does changing saturated fats to polyunsaturated fats in a person's diet **affect their risk of death**?

2.4.1 Reducing Serum Cholesterol

First, we want to test whether the unsaturated fat diet changes serum cholesterol levels. To do so, we will need the `serum_cholesterol` table. Remember that there are two unique values in the 'Condition' column: 'Diet' and 'Control'.

[115]:
```
serum_cholesterol
```

[115]:
```
Condition | Change in Serum Cholesterol
Diet      | -8.36662
Diet      | -23.6885
Diet      | -28.985
Diet      | -10.9341
Diet      | -17.9041
```

```
Diet      | -11.7145
Diet      | -13.6215
Diet      | -2.2387
Diet      | -2.03579
Diet      | -13.5746
... (2345 rows omitted)
```

Question 4.1. State precisely a null hypothesis and an alternative hypothesis which can help us determine if the unsaturated fat diet *decreases* serum cholesterol levels as compared to the control diet.

Null Hypothesis: unsaturated fat diet does not affect serum cholesterol levels (any difference is due to chance)

Alternative Hypothesis: unsaturated fat diet affects serum cholesterol levels

In order to differentiate between our two hypotheses above, we consider the difference in the average of the percentage changes between the control group and the diet group.

Question 4.2. Do larger values of the test statistic point towards the null hypothesis or the alternative hypothesis? Assign `larger_chol_stat` to either 1 if it's the null, or 2 if it's the alternative.

[117]:
```
larger_chol_stat = 1
```

Question 4.3. Define a function `compute_chol_test_statistic` which takes in a table just like `serum_cholesterol` and returns the test statistic of the given data. Remember that the "Change in Serum Cholesterol" column in the provided `tbl` for `compute_chol_test_statistic` will already have % changes.

[126]:
```
def compute_chol_test_statistic(tbl):
    grouped_chol = tbl.group('Condition', np.mean).column("Change in Serum␣
↪Cholesterol mean")
    percent_change_diet_chol = grouped_chol.item(1)
    percent_change_control_chol = grouped_chol.item(0)
    return abs(percent_change_diet_chol - percent_change_control_chol)
```

Question 4.4. Assign `chol_observed_statistic` to the value of the test statistic on the observed data.

[127]:
```
chol_observed_statistic = compute_chol_test_statistic(serum_cholesterol)
chol_observed_statistic
```

[127]:
```
12.829344627886611
```

Question 4.5. The next step in our hypothesis test is to simulate what we might observe if the null hypothesis were true. Describe the steps needed to simulate the test statistic under the null hypothesis. Then, write a function to simulate one value of the statistic under the null hypothesis.

To stimulate the test statistic under the null hypothesis, we will shuffle the values of 'Change in Serum Cholesterol' and use them to simulate statistics so as to determine if the observed statistic is extreme.

[130]:
```
def simulate_chol_change_null():
    shuffled_chol = serum_cholesterol.sample(with_replacement=False).
↪column('Change in Serum Cholesterol')
    sim_table_chol = serum_cholesterol.with_column('Change in Serum␣
↪Cholesterol', shuffled_chol)
    return compute_chol_test_statistic(sim_table_chol)
```

[131]:
```
# Run this cell to check that your function works.
simulate_chol_change_null()
```

[131]:
```
0.2545044027851473
```

Question 4.6. Simulate 1000 values of the test statistic by simulating taking a sample under the null hypothesis multiple times and assign this collection of test statistics to `chol_simulated_stats`. Put the test statistics into a one column table with 1000 rows called `chol_simulated_table`.

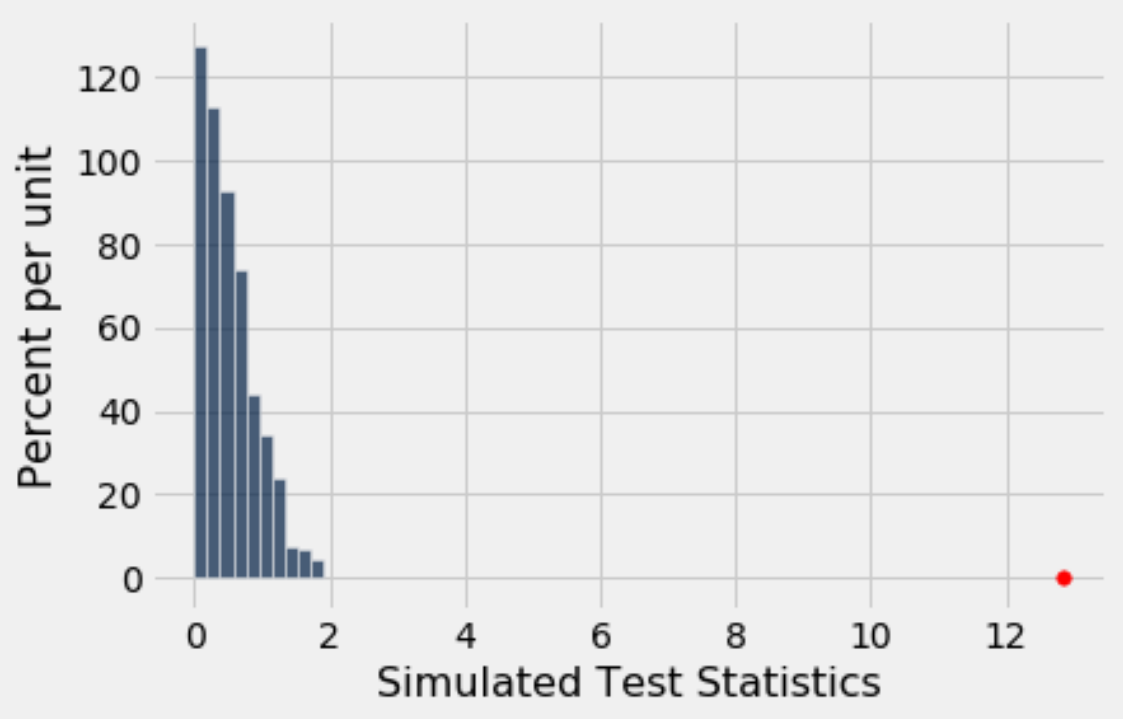
Note: Your code might take a couple of minutes to run.

[136]:
```
chol_simulated_stats = make_array()

for i in range(1500):
    sim_stat = simulate_chol_change_null()
    chol_simulated_stats = np.append(chol_simulated_stats, sim_stat)


chol_simulated_table = Table().with_column('Simulated Test Statistics',␣
 ↪chol_simulated_stats)
```

The following line plots the histogram of the simulated test statistics, as well as a point for the observed test statistic. Make sure to run it, as it will be graded.

[137]:
```
chol_simulated_table.hist()
plots.scatter(chol_observed_statistic, 0, color='red', s=30)
```

[137]:
```
<matplotlib.collections.PathCollection at 0x7f52b415d588>
```



Question 4.7. Without calculating any p-values, can we conclude from the test that the change in diet **causes** a larger percentage difference in serum cholesterol levels over time? Explain your answer.

No; without calculating any p-values, we cannot conclude from the test that an polyunsaturated fat diet causes a significant effect on serum cholesterol levels over time. If we do this, we are accepting the alternative hypothesis. When we rejecting the null hypothesis, it does not mean that the alternative hypothesis is true. The test could be indicating that an polyunsaturated fat diet causes a significant effect on serum cholesterol levels but does not prove that.

Question 4.8. Assign `cholesterol_conclusion` to 1, 2, or 3, where the number chosen corresponds to the conclusion that we can make from this study.

1. The results of this analysis indicate that changing saturated fats to polyunsaturated fats in a person's diet decreases their serum cholesterol levels.

2. The results of this analysis indicate that changing saturated fats to polyunsaturated fats in a person's diet does not decrease their serum cholesterol levels.

3. The results of this analysis do not allow us to draw any conclusions about the effect of changing saturated fats to polyunsaturated fats in a person's diet on their serum cholesterol levels.

[138]:
```
cholesterol_conclusion = 1
cholesterol_conclusion
```

[138]:
```
1
```

2.4.2 Reducing Death Rates

In the previous section, we made a decision on whether dietary change affects the change in serum cholesterol levels. We have not yet, however, explored how the change in diet affects death rates among the subjects. To explore this, we move our attention to the `minnesota_data` table.

[139]:
```
minnesota_data
```

[139]:
```
Age  | Condition | Died  | Participated
0-34 | Diet      | True  | True
0-34 | Diet      | True  | True
0-34 | Diet      | True  | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
0-34 | Diet      | False | True
... (9413 rows omitted)
```

Question 4.9. Set up a null hypothesis and an alternative hypothesis that we can use to answer whether or not the unsaturated fat diet causes different rates of death in the two groups.

Null Hypothesis: *The unsaturated fat diet does not cause a decreased rate of death in the two groups*

Alternative Hypothesis: *The unsaturated fat diet causes a decreased rate of death in the two groups*

Question 4.10. Create a table named `summed_mn_data`, with three columns and two rows. The three columns should be "Condition", "Died sum", and "Participated sum". There should be one row for the diet group and one row for the control group, and each row should encode the total number of people who participated in that group and the total number of people who died in that group.

[140]:
```
summed_mn_data = minnesota_data.group('Condition', sum).drop('Age sum')
summed_mn_data
```

[140]:
```
Condition | Died sum | Participated sum
Control   | 222      | 4738
Diet      | 245      | 4685
```

Question 4.11. In thinking of a test statistic, one researcher decides that the absolute difference in the number of people who died in the control group and the number of people who died in the diet group is a reasonable test statistic. Give one **specific** reason why this test statistic will not work.

This test statistic will not work because there are more participants in the control group than in the diet group

To combat the problem above, we instead decide to use the the absolute difference in hazard rates between the two groups as our test statistic. The *hazard rate* is defined as the proportion of people who died in a specific group out of the total number who participated in the study from that group.

Question 4.12. Define a new table `summed_mn_hazard_data` that contains the columns of `summed_mn_data` along with an additional column, `Hazard Rate`, that contains the hazard rates for each condition.

[149]:
```
summed_mn_hazard_data = summed_mn_data.with_column('Hazard Rate',summed_mn_data.
↪column(1)/summed_mn_data.column(2))
summed_mn_hazard_data
```

[149]:
```
Condition | Died sum | Participated sum | Hazard Rate
Control   | 222      | 4738             | 0.0468552
Diet      | 245      | 4685             | 0.0522946
```

Question 4.13. Define a function `compute_hazard_difference` which takes in a table like `summed_mn_hazard_data` and returns the absolute difference between the hazard rates of the control group and the diet group. Use it to get the observed test statistic and assign it to `death_rate_observed_statistic`.

[150]:
```
def compute_hazard_difference(tbl):
    return tbl.column('Hazard Rate').item(1) - tbl.column('Hazard Rate').item(0)

death_rate_observed_statistic = compute_hazard_difference(summed_mn_hazard_data)
death_rate_observed_statistic
```

[150]:
```
0.005439343927004493
```

Question 4.14. We are now in a position to run a hypothesis test to help differentiate between our two hypothesis using our data. Define a function `test` which takes in a table like `minnesota_data`. It simulates samples and calculates the rate differences for these samples under the null hypothesis 500 times, and uses them to return a P-Value with respect to our observed data. Note that your function should use the values in `t`, and should not refer to `minnesota_table`!

Hint: This is a long, involved problem. Start by outlining the steps you'll need to execute in this function and address each separately. Small steps and comments will be very helpful. You've already written a lot of key steps!

Note: Your code might take a long time to run.

[153]:
```
def test(t):
    diffs = make_array()
    repetitions = 1000

    for i in range(repetitions):
        shuffled_condition = t.sample(with_replacement=False).
↪column('Condition')
        sim_table_condition = t.with_column('Shuffled Condition',␣
↪shuffled_condition).drop('Condition')
        summed_t_data = sim_table_condition.group('Shuffled Condition', sum).
↪drop('Age sum')
        summed_t_hazard_data = summed_t_data.with_column('Hazard Rate',␣
↪summed_t_data.column(1)/summed_t_data.column(2))
        haz_diff = compute_hazard_difference(summed_t_hazard_data)
        diffs = np.append(diffs, haz_diff)
    return np.count_nonzero(diffs <= death_rate_observed_statistic)/repetitions

our_p_value = test(minnesota_data)
our_p_value
```

[153]:
```
0.898
```

Question 4.15. Using the P-Value above, what can we conclude about if the change in diet causes a difference in death rate? Assume a normal p-value cutoff of .05.

The p-value is greater than 0.05, therefore, we do not reject the null hypothesis. We can conclude that the unsaturated fat diet does not cause a decreased rate of death.

2.5 Part 5: Conclusion

We've almost made it to the end of this analysis. You, as an investigative data scientist, have explored the world's leading causes of death, identified the largest cause of death known to us in the last century, and looked at one of the most important data sets that explains what leads to that cause of death. We've recreated Dr. Frantz's data, run our own experiments, and examined important external factors. It's now time to reflect on what we've discovered.

Question 5.1. In about 3-5 sentences, explain what you have learned throughout this project. Does replacing saturated fats with unsaturated fats cause a change in serum cholesterol? Does it cause a different death rate? What other factors are important to consider?

This project helped me learn even further how to test hypotheses. I have also learned not to mistake an association for causality. Replacing saturated fats with unsaturated fats causes a change in serum cholesterol. However, it does not cause a difference in death rate. There are other factors to consider such as one's family medical history, gender and age, and perhaps general weight.

Submission: Congratulations! You have completed your own large scale case study into cause and effect surrounding one of the world's deadliest killers: cardiovascular disease. Your investigation you has taken you through two important data sets and across decades of medical research.

Time to submit. Once you're finished, submit your assignment as a .ipynb (Jupyter Notebook) and .pdf (download as .html, then print to save as a .pdf) on the class Canvas site.

2.5.1 Further reading

If you're interested in learning more, you can check out these articles:

- Origin story of the Framingham Heart Study
- Recent paper about Minnesota findings
- National Diet-Heart Study initial report
- National Diet-Heart Study final report