Name:_____ Date:_____

Practicum
TB or Not TB

Experiments in the medical field that involve new treatments (new medications) are called clinical trials. You have received a dataset that shows the results from Sir Austin Bradford Hill's first randomized study in 1948 examining the effects of the antibiotic Streptomycin on 107 tuberculosis patients. You and a partner will use this dataset to find out if Streptomycin is an effective treatment for tuberculosis.

A short article about tuberculosis facts can be found at:

https://www.cdc.gov/tb/communication-resources/tuberculosis-fact-sheet.html

Since this is an experiment, answer the following questions below. You may need to research the answer to some of the questions.

a. What is the research question?
b. Who are the subjects that participated in the experiment?
c. What is the treatment?
d. Who is in the treatment group?
e. Who is in the control group?
f. How were the subjects assigned to each group?
g. What population is this experiment representative of?
h. What is the variable that we will be measuring?
i. What is the outcome of this experiment?

To answer your research question, you and a partner will compare the outcome of the data with the outcomes given by a chance model (in which Streptomycin has no effect on TB).

1. First, scrape the data. Refer to the web scraping lab (Lab 3E) or copy/paste the code below into the console.

```
library(XML)
tb_url <- "https://labs.thinkdataed.org/extras/webdata/tb.html"
tb <- readHTMLTable(tb_url, which = 1)
names(tb) <- c("treatment", "outcome")
```

   Note: Alternatively, you can upload and import the CSV file (tb.csv) of the data into your RStudio project.

2. Second, determine the percentages of subjects in the study that died and the percentages of the subjects that recovered for each group.
3. Third, assuming that the treatment had no effect, use the data to:
   a. Calculate the percentage of people with tuberculosis we would expect to die.
   b. Use the *expected* percentage from (a), above, to calculate the number of people we expect to die from the treatment group.
   c. Compare the outcome from (b), the number of people we expected to die, to the number of people from the treatment group who *actually* died.
4. Then, if we assume that the outcome does not depend on the treatment, design and complete an appropriate simulation in RStudio using a chance model to replicate Sir Hill's study:
   a. Shuffle the treatment and control labels 300 times; each time, calculate the percentage of treatment patients who "died". Plot the distribution of the 300 percentages. Refer to the simulation labs if you need to recall how to create a simulation.
   b. Use the results from the chance model (shuffling) to determine whether (i.) or (ii.) below is the most reasonable explanation for the actual data in Sir Hill's study and state why:
      i. Streptomycin is a much better treatment for tuberculosis than bed rest. So, the outcome depends on the treatment.

Practicum
TB or Not TB

ii. The actual difference between treatments is due to chance; Streptomycin may not be effective on tuberculosis. So, it is possible that treatment and outcome are independent.

5. Can we say that Streptomycin ***causes*** the recovery of tuberculosis patients? Explain your answer.

Create a 4-5 slide, 5-minute presentation that shows your results. Be sure to include a detailed explanation of how you and your partner decided to conduct your simulation. Each person must participate in the presentation. In addition to the presentation, submit a 2–4-page, double-spaced summary of your analysis.