IDS Unit 3 – Practicum
TB or Not TB
Sample Solution

Preliminaries/ Question 1

First, we load the XML package and the TB data:

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

Question 2

Determine the percentages of subjects in the study that died and the percentages of the subjects that recovered for each group.

We start by creating 2 two-way frequency tables, one using percentages and the other using counts.

Note: The order of the variables matters. In this case, treatment is considered our x variable and the outcome is considered our y variable. So, we order them as such.

```
table1 <- tally(outcome~treatment, data=tb, format="percent")
table1
```

```
##      treatment
##  outcome          control      streptomycin
##      died         26.923       7.273
##      recovered    73.077       92.727
```

```
table2 <- tally(outcome~treatment, data=tb, format="count", margins=TRUE)
table2
```

```
##      treatment
##  outcome          control      streptomycin
##      died         14           4
##      recovered    38           51
##      Total        52           55
```

About 26.9231% of the control group died and 73.0769% recovered. For the treatment group, 7.2727% died and 92.7273% recovered.

Question 3a

Assuming that the treatment had no effect, use the data to:

a. Calculate the percentage of people with tuberculosis we would expect to die.

Assuming the medicine had no effect, we can estimate the percentage of people suffering from tuberculosis we would expect to die by dividing the number of people in the study that died by the total number of people in the study. We create a frequency table of died versus recovered to get these values.

```
table3 <- tally(~outcome, data=tb, margins=TRUE)
table3
```

```
##      outcome
##      died   recovered    Total
##      18     89           107
```

Doing this, we estimate the expected percentage of people to die to be 16.8224%.

IDS Unit 3 – Practicum
TB or Not TB
Sample Solution

Question 3b

Assuming that the treatment had no effect, use the data to:

b. Use the expected percentage from (a), above, to calculate the number of people we expect to die from the treatment group.

Applying this number to the 55 people in our treatment group, we calculate the number of people we expect to die in the treatment group to be 9.2523 people (55 • 0.168224). Again, this is under the assumption that the medicine had absolutely no effect whatsoever.

Question 3c

Assuming that the treatment had no effect, use the data to:

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.

So, we would expect approximately 9 people to die in the treatment group but in actuality we had 4 people who died in the treatment group.

Question 4a

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:

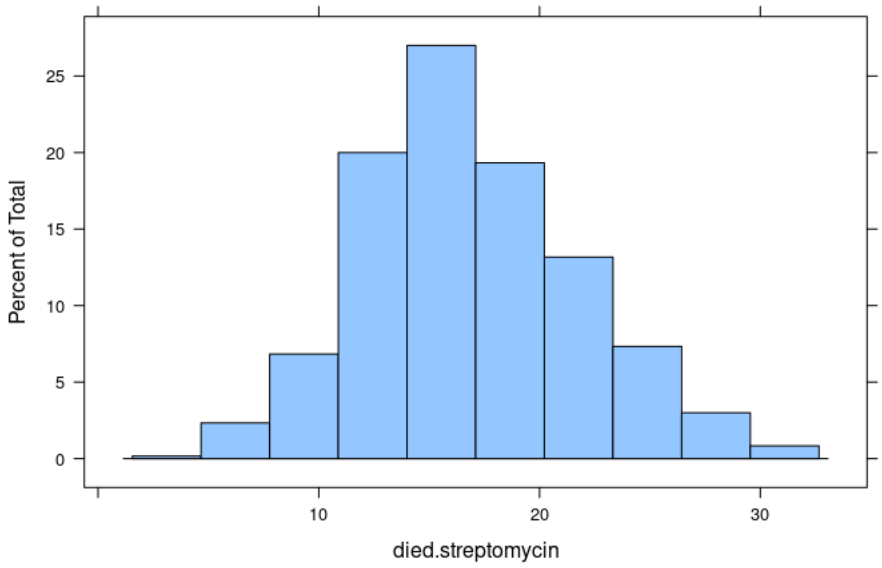
a. Shuffle the treatment and control labels 600 times; each time, calculate the percentage of treatment patients who "died". Plot the distribution of the 600 percentages. Refer to the simulation labs if you need to recall how to create a simulation.

To create our chance model, we run the following (`set.seed()` is not required but may help with reproducibility):

```
set.seed(297)
shfl_tb <- do(600) * tally(outcome~treatment,
                data=resample(tb, shuffled="treatment"), format="percent")
```

We can then make a histogram for the number of people who died using our chance model:

```
histogram(~died.streptomycin, data=shfl_tb, type="percent")
```



IDS Unit 3 – Practicum
TB or Not TB
Sample Solution

Question 4b

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:

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.

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.

Based on the histogram, we would say that (i.) is correct because the actual percentage of people who died (7.273) doesn't occur very often by chance alone. We can confirm this by calculating the probability.

```
mean_shfl_tb <- mean(~died.streptomycin, data=shfl_tb)
mean_shfl_tb

sd_shfl_tb <- sd(~died.streptomycin, data=shfl_tb)
sd_shfl_tb

pnorm(7.273, mean=mean_shfl_tb, sd=sd_shfl_tb)
[1] 0.02721505
```

The probability of getting 7.273% or lower deaths while on the treatment is 2.72%, which confirms our conclusion from the histogram. This leads us to believe that Streptomycin is having some sort of effect which is helping the patients recover. Thus, we conclude that Streptomycin is a much better treatment for TB than bed rest.

Conclusion/ Question 5

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

Since the patients were randomly assigned into treatment and control groups and since, holding all else equal, the patients receiving Streptomycin were more likely recover than those only receiving bed rest, we can conclude that Streptomycin directly helps TB patients recover.