IDS Unit 3 – End of Unit Project
TB or Not TB
Sample Solution

Preliminaries

First we load the XML package and the tb data:

```
library(XML)
tb_url <- "http://web.ohmage.org/mobilize/resources/ids/data/tb.html"
tables <- readHTMLTable(tb_url)
tb <- tables[[1]]
names(tb) <- c("treatment","outcome")
```

Question 1

First, 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 with using percentages and the other using counts.

NOTE: The order of the variables matters. In this case, treatment is considered our X variable and 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 2

Second, assuming that the treatment had no effect, use the data to estimate the percentage of people with tuberculosis whom we would expect to die. Then, using the expected percentage, calculate the number of people we expect to die from the treatment group. Compare this number to the number of people from the treatment group who actually died.

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

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. Again, this is under the assumption that the medicine had absolutely no effect what-soever.

Question 3

Third, 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 Hill's study:

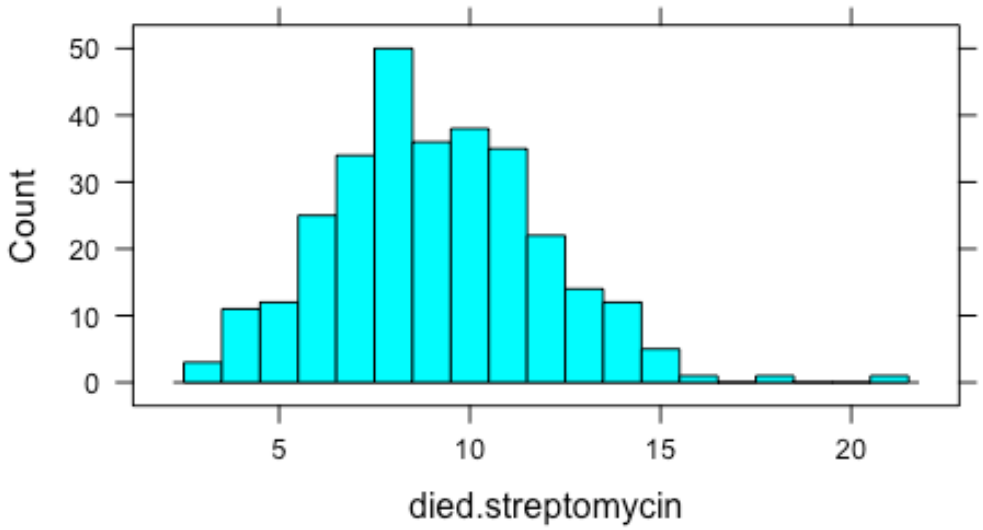
Shuffle the data 300 times and create a plot that shows your results.

To create our chance model, we run the following:

```
shfl_tb <- do(300) * tally(outcome~treatment, data=resample(tb, shuffled="
treatment"), format="count")
```

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

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



Use the results from the chance model (shuffling) to determine whether (1.) or (2.) below is the most reasonable explanation for the actual data in Hill's study:

1. *Streptomycin is a much better treatment for tuberculosis than bed rest. So the outcome depends on the treatment.*
2. *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.*

We conclude that (1.) is correct because the actual number of people who died (4) doesn't occur very often by chance alone. 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.

IDS Unit 3 – End of Unit Project
TB or Not TB
Sample Solution

Question 4

If we conclude that (1.) is the explanation, can we say that Streptomycin causes the recovery of tuberculosis patients? Explain.

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.