Stat 145 – Malaria Study

A small Malaria Study

Let's consider a small study of a new malaria vaccine called PfSPZ. In this study, 20 volunteer patients were randomized into one of two experiment groups:

- 14 patients received the new vaccine;
- 6 patients received a placebo vaccine (a vaccine with no active ingredient).

Nineteen weeks later, all 20 patients were exposed to a drug-sensitive malaria virus strain. Here are the results:

- 5 of the 14 treatment patients had malaria symptoms;
- all 6 of the patients in the placebo group had malaria symptoms.

Questions about this study

Appoint one person in your group to be the **scribe for the day**. This person should take notes on your answers in an R Markdown document (remember to start from the Stat 145 template) and email a PDF to `rpruim@calvin.edu` when class is over.

Make sure to put all of your names on the sheet.

You might find it handy to have your scribe share their screen so you can all see.

1. Is this an observational study or an experiment? Why?

2. Why were the six patients given a vaccine at all if it had no active ingredient anyway?

3. What does it mean for a malaria strain to be "drug-sensitive"? Why do you think the patients were given a *drug-sensitive* malaria virus strain?

4. What are the variables in this study? For each, say whether they are categorical or numerical.

Send a Teams message to professor Pruim to let him know that your group has finished the first 4 problems.

5. Create a 2-way table showing the results of this study. (Do this by hand, not in R.)

6. What proportion of the treatment group contracted malaria? (Again, do it by hand, not in R.)

7. What proportion of the placebo group contracted malaria? (Again, do it by hand, not in R.)

8. Now get R do answer questions 5-7. You can load the data with the following command:

```
Malaria <- read.csv('https://rpruim.github.io/s145/data/malaria.csv')
head(Malaria)    # see what the first few rows look like
```

Send a Teams message to professor Pruim to let him know that your group has finished the first 8 problems.

9. Since we have the data in R, let's make a bar chart, too.

10. Based on those two proportions, does it appear that the new treatment is effective?

11. Besides the effectiveness of the new vaccine, what could explain a difference in these proportions?

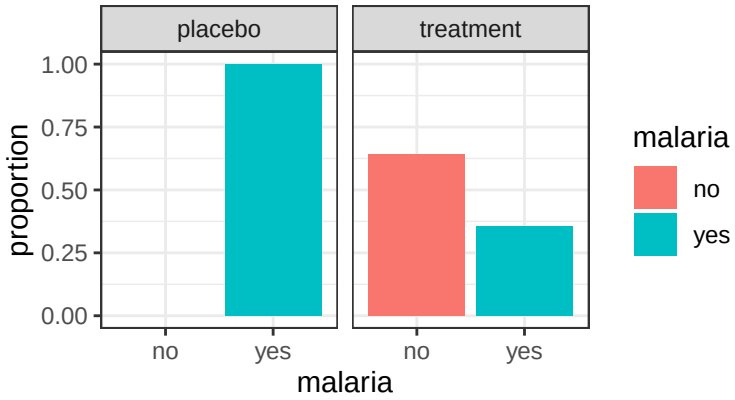
12. If you were asked to just whether the difference in these proportions was an indication that the new vaccine is effective, what might you do? (Be as specific as you can. You might find it useful to think back to the lady tasting tea example, but this situation is a little bit different, so think carefully.)

Send a Teams message to professor Pruim to let him know that your group has finished problem 12 and head back to the main class meeting.

Looking at the study data

Here's how we can get R to summarise the results of the study for us.

```
gf_props( ~ malaria | group, data = Malaria, fill = ~ malaria, position = "dodge")
```




```
tally( ~ malaria + group, data = Malaria, margins = TRUE)
```

```
##        group
## malaria placebo treatment Total
##   no          0         9     9
##   yes         6         5    11
##   Total       6        14    20
```

```
tally( malaria ~ group, data = Malaria, margins = TRUE, format = "prop")
```

```
##        group
## malaria   placebo treatment
##   no    0.0000000 0.6428571
##   yes   1.0000000 0.3571429
##   Total 1.0000000 1.0000000
```

```
diffprop( malaria ~ group, data = Malaria)
```

```
##  diffprop
## 0.6428571
```

A simulation

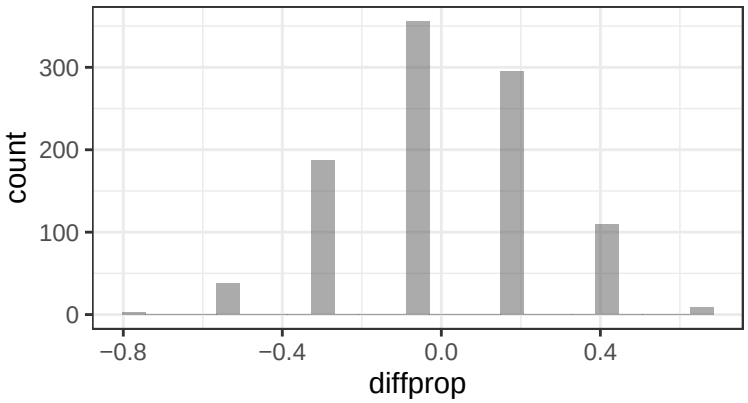
Two see whether the observed difference in proportions could be attributed to chance, we will first perform a physical simulation. Later, we will let R automate this for us (because that's much faster). But first we want to understand the process.

13. Make 20 cards. Each card represents a person. Label 11 of them with a sad face. Label the other 9 with a happy face. (Put your labels on one side only.)
    a. Why are there 20 cards?
    b. Why are there 11 sad faces?
    c. Why are there 9 happy faces?
14. Now shuffle the cards thoroughly and deal them into two piles – a pile with 14 cards and a pile with 6 cards.
    a. What do the two piles represent?
15. Create a two-way table for the cards you just dealt.

Face	large pile	small pile
Happy		
Sad		

16. Compute
    a. the proportion of sad faces in each pile,
    b. the difference between these two proportions.
17. Shuffle the cards and deal them out several more times. Each time calculate the difference in proportions and record the result.
    a. How many of your samples had a difference in proportion **at least as large** as what was observed in the actual study?
    b. How many of your samples had a difference in proportion **smaller than** what was observed in the actual study?
    c. We'll collect data from the whole class to see how unusual the difference of proportions observed in the study seems to be. (Next time we'll see how to get R to do this many more times.)

18. Here is a histogram and a table showing the difference in proportions for 1000 simulated studies, "shuffling the deck" each time, just as you were doing above. What conclusion would you draw based on looking at this histogram?



-0.7857	-0.5476	-0.3095	-0.0714	0.1667	0.4048	0.6429
3	38	188	356	296	110	9

Running the simulation in R

To create a histogram like the one above we need to mimic what we did with the cards:

- Create random data sets
- For each random data set, compute the difference in proportions
- Compare the original observed difference in proportions to the distribution of of these differences from our simulations

Let's start with the last step. The `diffprop()` function makes it easy to compute a difference in proportions:

```
diffprop(malaria ~ group, data = Malaria)
```

```
##  diffprop
## 0.6428571
```

Now let's think about the shuffling. Here's our original data (in a terser format):

```
## Rows: 20
## Columns: 2
## $ group   <fct> P, P, P, P, P, P, T, T, T, T, T, T, T, T, T, T, T, T, T, T
## $ malaria <fct> Y, Y, Y, Y, Y, Y, N, N, N, N, N, N, N, N, N, Y, Y, Y, Y, Y
```

And here are a few reshufflings of `malaria`:

```
## Rows: 20
## Columns: 2
## $ group   <fct> P, P, P, P, P, P, T, T, T, T, T, T, T, T, T, T, T, T, T, T
## $ malaria <fct> N, Y, Y, Y, N, Y, Y, N, N, Y, Y, N, N, Y, Y, Y, N, N, N, Y
```

```
## Rows: 20
## Columns: 2
## $ group   <fct> P, P, P, P, P, P, T, T, T, T, T, T, T, T, T, T, T, T, T, T
## $ malaria <fct> N, N, Y, Y, Y, Y, N, Y, N, N, Y, N, Y, N, Y, N, N, Y, Y, Y
```

```
## Rows: 20
## Columns: 2
```

```
## $ group   <fct> P, P, P, P, P, P, T, T, T, T, T, T, T, T, T, T, T, T, T, T
## $ malaria <fct> N, Y, Y, N, N, Y, Y, N, Y, N, N, Y, Y, Y, Y, Y, N, N, Y, N
```

All we really need is the difference in proportions for each reshuffling. Here's how we can get that.

```
diffprop( shuffle(malaria) ~ group, data = Malaria)
```

```
##  diffprop
## 0.4047619
```

```
diffprop( shuffle(malaria) ~ group, data = Malaria)
```

```
##  diffprop
## 0.1666667
```

```
diffprop( shuffle(malaria) ~ group, data = Malaria)
```

```
##  diffprop
## 0.4047619
```

Here's an even better way

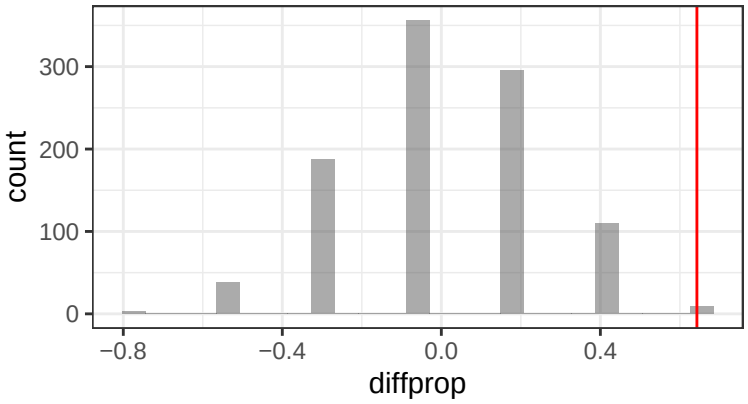
```
do(3) * diffprop( shuffle(malaria) ~ group, data = Malaria)
```

```
##      diffprop
## 1 -0.07142857
## 2 -0.30952381
## 3 -0.30952381
```

Now it is easy to do this lots of times

```
set.seed(123)
MalariaSims <- do(1000) * diffprop( shuffle(malaria) ~ group, data = Malaria)
gf_histogram(~diffprop, data = MalariaSims) %>%
  gf_vline(xintercept = 0.6428, color = "red")
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```



If we want a numerical summary, we can do one of these:

```
tally( ~ (diffprop >= 0.6428), data = MalariaSims)
```

```
## (diffprop >= 0.6428)
##  TRUE FALSE
##     9   991
```

```
tally( ~ (diffprop >= 0.6428), data = MalariaSims, format = "prop")
```

```
## (diffprop >= 0.6428)
##  TRUE FALSE
## 0.009 0.991
```

```
prop( ~ (diffprop >= 0.6428), data = MalariaSims)
```

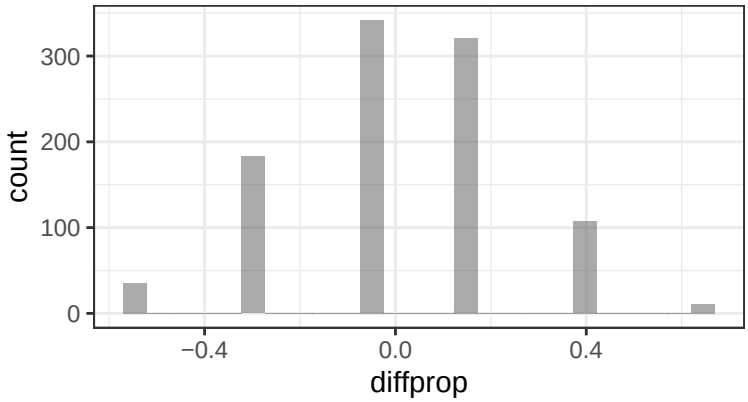
```
## prop_TRUE
##     0.009
```

In any case, we see that about 1% of random shufflings of malaria lead to a difference in proportions at least as large as what we saw in the original data. So is the effect due to random chance? It is possible, but not very likely. This provides evidence that the vaccine may be effective.

Notes:

1. `set.seed()` sets the seed for the randomization. If we use this function, then we will get the same random shuffling each time we run our R Markdown document. Otherwise, it will change a little it each time, which makes it harder to write about the results. If you change the number inside `set.seed()`, then you will get a different shuffle.
2. It is a good idea to build up gradually to your simulation.
    a. Calculate the desired statistic from the original dataset.
    b. Calculate the desired statistic from one shuffled dataset.
    c. `do()` that a few times to make sure things seem to be working.
    d. Finally, `do()` it a lot of times (1000 or 2000 usually suffices for this class. For a research paper, you might do it 10,000 times or more.) Save the results.
    e. Use numerical and/or graphical summaries to see how unusual the result from the original dataset would be in a randomly simulated dataset.
3. We could shuffle `group` instead of `malaria`. (We could even shuffle both, I guess.) Either way, the shuffling randomly matches up the treatment group to the outcome, exactly as would be the case if there were no association between the two.

```
MoreMalariaSims <-
  do(1000) * diffprop( malaria ~ shuffle(group), data = Malaria)
gf_histogram(~diffprop, data = MoreMalariaSims)
```



4. This sort of study is usually expressed in terms of two hypotheses. We have two hypotheses here:
    - Null Hypothesis (H_0): Treatment group and malaria outcome are independent, there is no association between the two **in the population**.
    - Alternaitve Hypothesis (H_a): Treatment group and malaria outcome are not independent, there is an association between the two **in the population**.

We run our simulation assuming the null hypothesis is true. If the results from the actual study would be quite unusual when H_0 is true, that provides evidence in favor of the alternative.

Note: A **hypothesis** is simply a statement that could be true or false. A **statistical hypothesis** is a hypothesis about a parameter or parameters. (In this case our hypotheses are about the difference in proportions of people who get malaria if they do or do not get the vaccine.) The hypotheses are about the entire population, not our data set.