Lab 9: Malaria

Two-way Chi-square Test

Dr. Robinson's Key

An article that appeared in the journal Lancet in May of 2021 (Datoo, et al.) described a study that investigated a potential vaccine that might protect children against malaria. Researchers recruited children between the ages of 5 and 17 months in Burkina Faso, a country in western Africa, as participants. The children were randomly assigned into one of three groups: one group received a large dose of the potential vaccine, another group received a small dose, and a third group received a placebo. Researchers observed the children for the next 18 months, keeping track of whether or not the child developed malaria. Researchers hoped, of course, that children who received a vaccine would be less likely to develop malaria than children who received a placebo.

Setup

```
malaria_data <- read_csv("data/malaria_data.csv")
malaria_data
```

```
# A tibble: 439 x 3
      ID Vaccine   Malaria
   <dbl> <chr>     <chr>
 1     1 Low Dose  Did Not Develop Malaria
 2     2 High Dose Did Not Develop Malaria
 3     3 High Dose Did Not Develop Malaria
 4     4 Low Dose  Developed Malaria
 5     5 Placebo   Developed Malaria
 6     6 High Dose Developed Malaria
 7     7 Placebo   Developed Malaria
 8     8 Placebo   Developed Malaria
 9     9 Low Dose  Did Not Develop Malaria
10    10 Placebo   Did Not Develop Malaria
# ... with 429 more rows
```

1. What is the observational unit for the study?

A child between the age of 5 and 17 in Burkina Faso.

2. Which is the explanatory variable and which is the response variable? Specify the variable types and levels/units

- Explanatory: Vaccine: High dose, Low dose, Placebo
- Response: Malaria: Developed Malaria, Did Not Develop Malaria

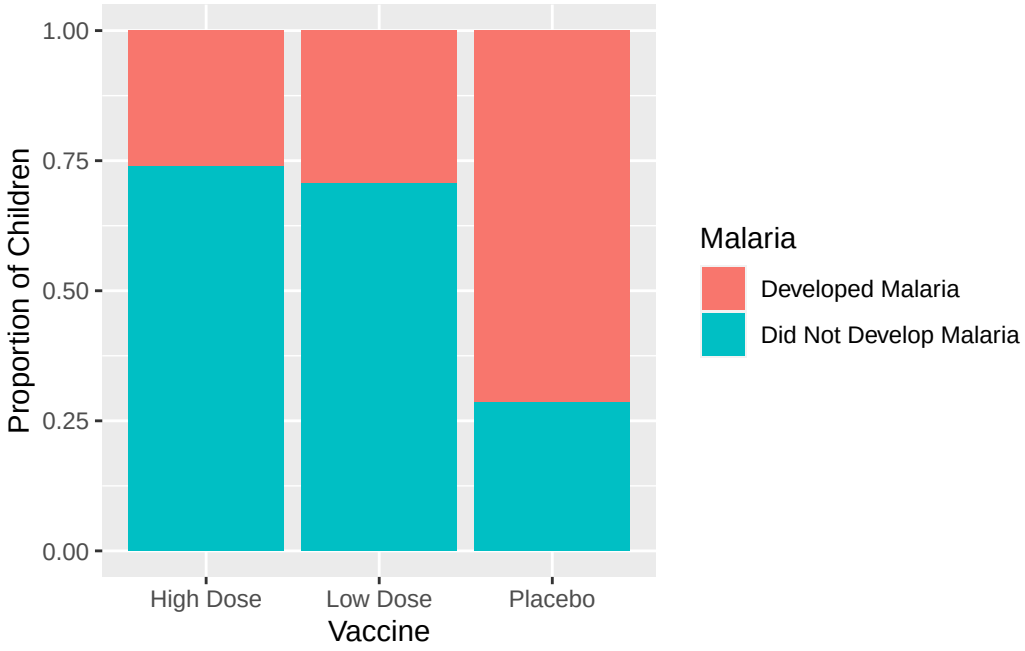
3. Should you do a chi-square test of independence or test of homogeneity? Justify how you know.

We should conduct a chi-square test of homogeneity since the vaccine level was randomly assigned to children.

Visualize & Summarize the Data

4. Fill in the code below to make a filled bar plot of the data. What do you see in your filled-bar plot?

```
ggplot(data = malaria_data,
       mapping = aes(x = Vaccine, fill = Malaria)) +
  geom_bar(position = "fill") +
  labs(x = "Vaccine",
       y = "Proportion of Children")
```



5. Why do we want to make a filled bar plot over a stacked or dodged bar plot?

Since we are conducting a chi-square test of homogeneity, we can compare the proportions across groups. The filled bar plot displaysthe proportions instead of the counts.

6. Finish the code below to make a two-way contingency table. Note: we typically want our explanatory variable to be indicated in the columns.

```
malaria_data %>%
   count(Vaccine, Malaria)%>%
   pivot_wider(names_from = Vaccine,
               values_from = n) %>%
   adorn_totals(where = c("row", "col"))
```

```
                Malaria High Dose Low Dose Placebo Total
      Developed Malaria        38       43     105   186
Did Not Develop Malaria       108      103      42   253
                  Total       146      146     147   439
```

7. What proportion of children who received the high dose vaccine contracted malaria?

0.26%

8. What proportion of children who received the low dose vaccine contracted malaria?

0.295%

9. What proportion of children who received the placebo contracted malaria?

0.714%

Theory-based Chi-square

10. Write the null and alternative hypothesis for this study in words.

- Null: The true proportion of children who contracted malaria is equal for all vaccine levels.
- Alternative: At least one vaccine level has a different proportiono of children who contracted malaria.

Recall, in order for the χ^2 distribution to be a good approximation of the true sampling distribution, we need to verify two conditions:

- The observations are independent
- We have a "large enough" sample size
  - This is checked by verifying there are at least 5 expected counts in each cell

11. Is the independent observation condition met? Justify your answer.

Yes, knowing one child's malaria status does not tell us information about another child's malaria status.

The equation for calculating expected counts is:

$$\frac{\text{row } i \text{ total count} \times \text{column } j \text{ total count}}{\text{total count}}$$

Our table is only a 2 x 3 table, but what if you had a 6 x 8 or worse, 20 x 42 table? Checking each cell's expected count would be very tedious. We only need to check the cell which will have the smallest expected count.

12. Which row in your two-way contingency table from #6 has the smallest total count?

The 'Developed Malaria' row has the smallest total count at 186.

13. Which column in your two-way contingency table from #6 has the smallest total count?

The 'High Dose' and 'Low Dose' columns have the smallest total count at 146.

14. Using the equation above calculate the expected count for the cell in the row and column you specified in #9 and #10.

The smallest expected count for Developing Malaria in the High Dose and Low Dose vaccine groups is $\frac{186 \times 146}{439} = 61.859$

15. Is the "large enough" sample size condition met?

Yes, the smallest expected count (61.859) is at least 5; therefore, we have a 'large enough' sample size.

16. Can we use the χ^2 distribution to approximate the true sampling distribution? What conditions did we have to check?

Yes, since independence and 'large enough' sample size are both met, we can use the χ^2 distribution to approximate the true sampling distribution.

17. Fill in the code below to perform a theory based Chi-square test.

```
chisq_test(x = malaria_data,
           response = Malaria,
           explanatory = Vaccine)
```

```
# A tibble: 1 x 3
  statistic chisq_df  p_value
      <dbl>    <int>    <dbl>
1      76.8        2 2.12e-17
```

18. What conclusion would you reach based on your results? *Make sure to address (1) Chi-square test statistic and associated degrees of freedom, (2) p-value, (3)* α threshold, (4) your decision about the null hypothesis, (5) your conclusion in context of the data, and (6) the scope of inference.

With a $X_2^2 = 76.78$ and p-value of less than 0.0001, we reject the null hypothesis with a significance level of 0.05. We have sufficient evidence to conclude at least one vaccine group has a different true proportion of developing Malaria. Given the vaccine treatment was randomly assigned to children, researchers can say the vaccine causes a difference in the true proportion of developing Malaria.

Simulated Chi-square

What if our conditions had not been met? We would have needed to use a simulation based approach. Let's walk through what this would look like.

19. First, we need to calculate the observed chi-square test statistic from our data. We did this by hand during the activities, but it can be tedious so let's make R do it for us. Fill in the code below to calculate your observed chi-square test statistic.

```
obs_xsq <- malaria_data %>%
  specify(response = Malaria,
          explanatory = Vaccine) %>%
  calculate(stat = "Chisq")

obs_xsq
```

```
Response: Malaria (factor)
Explanatory: Vaccine (factor)
# A tibble: 1 x 1
   stat
  <dbl>
1  76.8
```

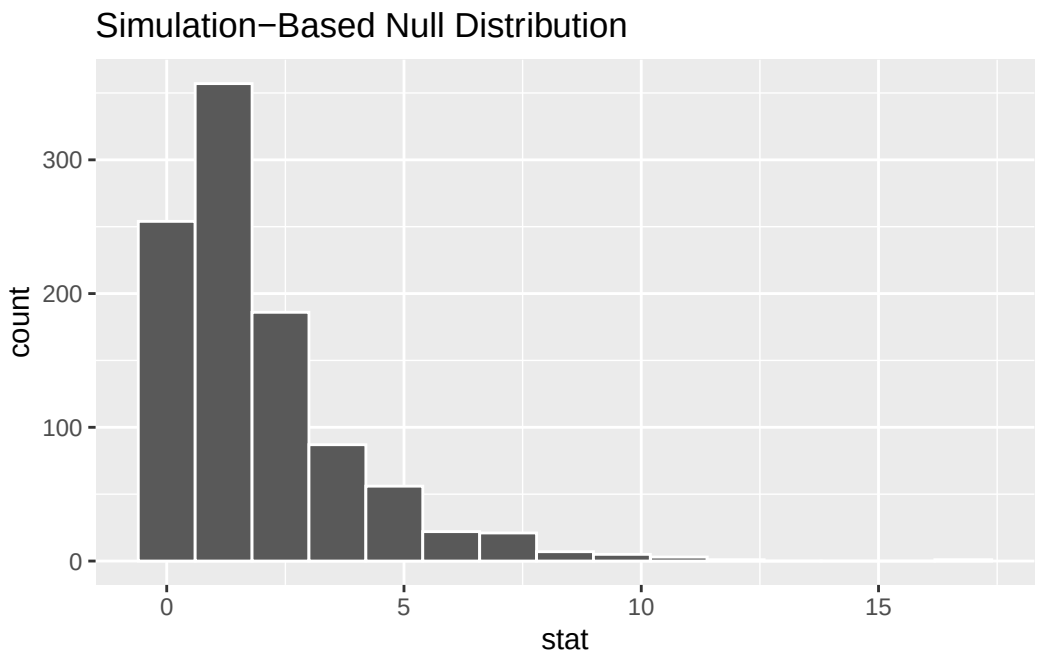
20. What value of the observed chi-square test statistic did you calculate above? Where have we previously seen this?

The observed chi-square test statistic is c('X-squared' = 76.78). This is the same value outputted in our chi-square test from question #16.

21. Now we need to generate what the sampling distribution would look like if the null were true (aka null distribution of our chi-square statistics). Fill in the code below to generate and visualize the null distribution.

```
null_dist <- malaria_data %>%
  specify(response = Malaria,
          explanatory = Vaccine) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")

visualize(data = null_dist,
          method = "simulation")
```



22. Once we have our null distribution, we can use this to calculate our simulated p-value.

```
get_pvalue(x = null_dist,
           obs_stat = obs_xsq,
           direction = "greater")
```

```
Warning: Please be cautious in reporting a p-value of 0. This result is an
approximation based on the number of `reps` chosen in the `generate()` step. See
`?get_p_value()` for more information.
```

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```

23. What conclusion would you reach with the simulated chi-square test? Does this differ from your answer in #17?

With a p-value of approximately 0, we would reject the null hypothesis with a significance level of 0.05. We have sufficient evidence to conclude at least one vaccine group has a different true proportion of developing Malaria. This is the same conclusion as in #17 since our conditions were met.