Lab 9: Malaria

Two-way Chi-square Test

Present Group Member's Names Go Here!

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?

answer goes here…

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

- Explanatory:
- Response:

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

answer goes here…

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 = __________,
       mapping = aes(x = _________, fill = _______)) +
  geom_bar(position = "______") +
  labs(x = "____________",
       y = "____________")
```

```
## Error: <text>:1:15: unexpected input
## 1: ggplot(data = _
##                   ^
```

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

   answer goes here...

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(_____________, _____________)%>%
   pivot_wider(names_from = _______________,
               values_from = n) %>%
   adorn_totals(where = c("row", "col"))
```

```
## Error: <text>:2:10: unexpected input
## 1: malaria_data %>%
## 2:    count(_
##             ^
```

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

answer goes here...

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

answer goes here...

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

answer goes here...

Theory-based Chi-square

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

- Null:
- Alternative:

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.

answer goes here…

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?

    answer goes here…

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

    answer goes here…

14. Using the equation above calculate the expected count for the cell in the row and column you specified in #10 and #11.

answer goes here…

15. Is the “large enough” sample size condition met?

answer goes here…

16. Can we use the χ^2 distribution to approximate the true sampling distribution? Which conditions were necessary to have been checked?

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

```
chisq_test(x = malaria_data,
           response = ________________,
           explanatory = ________________)
```

```
## Error: <text>:2:23: unexpected input
## 1: chisq_test(x = malaria_data,
## 2:            response = _
##                          ^
```

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, and (5) your conclusion in context of the data.

answer goes here…give me a nice paragraph!

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 = _____________________,
          explanatory = __________________) %>%
  calculate(stat = "_______")

obs_xsq
```

```
## Error: <text>:2:22: unexpected input
## 1: obs_xsq <- malaria_data %>%
## 2:   specify(response = _
##                         ^
```

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

answer goes here…

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 = "______________") %>%
  generate(reps = 1000, type = "__________") %>%
  calculate(stat = "Chisq")
```

```
## Error: The ‘type‘ argument should be one of "bootstrap", "permute", or "draw". See ‘?generate‘ for mo
```

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

```
## Error in visualize(data = null_dist, method = "simulation"): object 'null_dist' not found
```

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 = "__________")
```

```
## Error in get_pvalue(x = null_dist, obs_stat = obs_xsq, direction = "__________"): object 'null_dist'
```

23. What conclusion would you reach with the simulated chi-square test? Does this differ from your answer in #18?

answer goes here…