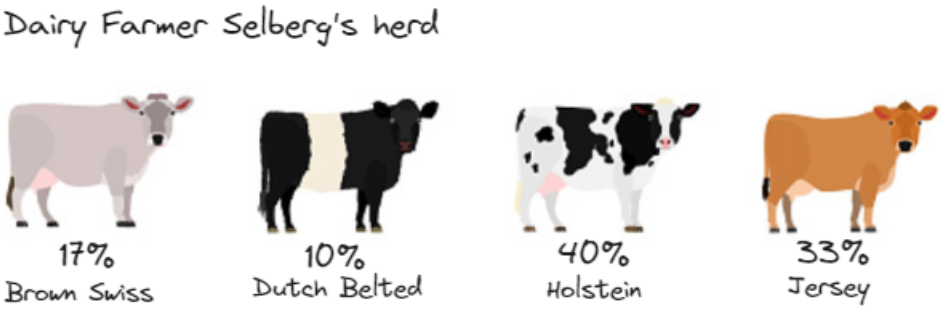
Homework 3: Which cow breeds have mastitis?

Inference for a Categorical Variable with More than Two Categories

Mastitis is a common and costly condition in dairy cows and heifers. It is an inflammation of the mammary gland tissue and udder, which can be caused by various factors, including bacterial infections, environmental conditions, and poor milking practices. Mastitis can lead to reduced milk production and quality. Identifying whether the breed pattern of cows with mastitis consistent with the breed pattern of cows in the herd and is crucial for dairy farmer Selberg.

The proportion of cows of each breed in the dairy farmer Selberg's herd is shown below.



Farmer Selberg has collected data (shown below) on the heifers and diary cows with mastitis from his herd.

```
library(tidyverse)                                    ①
mastitis <- read_csv("data/mastitis.csv")             ②
head(mastitis)                                        ③
```

① Load the `tidyverse` package.
② Read in the `mastitis` data set.
③ View the top 6 rows of the `mastitis` data set.

```
# A tibble: 6 x 5
  cow_id condition breed      age_months type
   <dbl> <chr>     <chr>           <dbl> <chr>
1      1 mastitis  BrownSwiss         39 dairy cow
2      2 mastitis  BrownSwiss         35 heifer
3      3 mastitis  BrownSwiss         48 dairy cow
4      4 mastitis  Holstein           32 heifer
5      5 mastitis  Jersey             32 heifer
6      6 mastitis  BrownSwiss         30 heifer
```

Research Question: Is there evidence to suggest that the breed pattern of cows with mastitis deviates from breed patterns in dairy farmer Selberg's herd?

1. Identify the variable of interest (and categories).

2. State the parameters and appropriate symbols *(Hint: there should be 4)*.

3. Write your null and alternative hypotheses.

Consider the following table. The first row of this table contains the Observed number of cows with mastitis. The second row contains the Expected number of cows with mastitis (under the null hypothesis) for each of the breeds.

	Brown Swiss	Dutch Belted	Holstein	Jersey	Total
Observed	39	6	62	31	138
Expected	23.46				

4. What does the Total value for the Observed row represent?

5. The value in the first row and second column is 23.46 (i.e. Expected count for the Brown Swiss breed). Explain where this number came from. What does this value represent?

6. Complete the empty cells in the second row to contain the expected count for each of the breeds.

7. Sketch a stacked bar plot of the Observed *counts* of mastitis versus Expected *counts* of mastitis for each breed.

8. Sketch a stacked bar plot of the Observed *proportions* of mastitis versus Expected *proportions* of mastitis for each breed (*Hint: you may want to create a second table containing the observed and expected proportions)*

9. Why is the Expected count for the Holstein's higher than the other breeds?

10. Suppose your friend computes the following percentages: Brown Swiss: $39/138 \approx 28\%$; Dutch Belted: $6/138 \approx 4\%$; Holstein: $62/138 \approx 45\%$; and Jersey: $31/138 \approx 23\%$. Your friend then makes the following statement: "There is enough evidence for the research question because these percentages are different from the breed percentages for the entire herd (i.e., Brown Swiss = 17%, Dutch Belted = 10%, Holstein = 40%, and Jersey = 33%)." Why is this statement statistically incorrect? Explain.

11. Calculate the Chi-square test statistic for our sample of cows *(Hint: it may be helpful to add additional rows to the table above).*

12. Conduct 1000 replications via the `Online Simulation Applets` > `Goodness of Fit` to create a simulated distribution of the test statistic under the assumption that cows with mastitis do not deviate from the overall breed patterns in dairy farmer Selberg's herd. *(Hint: you will need to change the hypothesized probabilities in the applet).*

    The breed data can be found at https://raw.githubusercontent.com/earobinson95/stat218-calpoly/main/02-homeworks/data/mastitis_breeds.csv

> **Warning**
>
> Watch the ordering of the categories when setting up your hypothesized probabilities.

    Paste/sketch the simulated null sampling distribution below.

13. Is our observed Chi-square test statistic consistent with results we would expect to see if cows with mastitis do not deviate from the overall breed patterns in dairy farmer Selberg's herd? Explain.

14. From your simulated distribution, compute the p-value, make a decision, and write a final conclusion for the original research question.

Recall that in practice, statisticians conduct a Goodness of Fit test using the Chi-square distribution (and code) rather than simulating the distribution.

12. Check the conditions for using the Chi-square distribution.

13. How many degrees of freedom are used to determine the shape of the Chi-square distribution for our scenario?

14. Select the correct code/output below to test the hypotheses. Explain your selection.

a.

```
library(infer)
chisq_test(mastitis,
           response = breed,
           p = c(0.28, 0.04, 0.45, 0.23)
           )
```

```
# A tibble: 1 x 3
  statistic chisq_df p_value
      <dbl>    <dbl>   <dbl>
1    0.0625        3   0.996
```

b.

```
library(infer)
chisq_test(mastitis,
           response = breed,
           p = c(0.25, 0.25, 0.25, 0.25)
           )
```

```
# A tibble: 1 x 3
  statistic chisq_df  p_value
      <dbl>    <dbl>    <dbl>
1      46.4        3 4.65e-10
```

c.

```
library(infer)
chisq_test(mastitis,
           response = breed,
           p = c(0.17, 0.10, 0.40, 0.33)
           )
```

```
# A tibble: 1 x 3
  statistic chisq_df  p_value
      <dbl>    <dbl>    <dbl>
1      20.2        3 0.000156
```

15. From the output above, find the p-value, make a decision, and write a final conclusion for the original research question.

Canvas Quiz

Make sure to complete the Homework Quiz on Canvas.