Chi-Square Test of Independence: Fatal Injuries in the Iliad

Learning outcomes

- Given a research question involving two categorical variables, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a chi-square test of independence for two categorical variables by:
  - Finding the expected counts
  - Calculating the X^2 statistic
- Interpret and evaluate a p-value for a chi-square test of independence for two categorical variables.
- Use mathematical conditions to determine whether simulation-based methods or theory-based methods should be used when obtaining a p-value.

Fatal Injuries in the Iliad

Homer's Iliad is an epic poem, compiled around 800 BCE, that describes several weeks of the last year of the 10-year siege of Troy (Ilion) by the Achaeans. The story centers on the rage of the great warrior Achilles. But it includes many details of injuries and outcomes, and is thus the oldest record of Greek medicine. The data report 146 recorded injuries for which both injury site and outcome are provided in the Iliad (Hutchinson, 2013).

For this activity we will focus on assessing if the location of an injury is associated with whether the injury was fatal.

Exploratory Data Analysis

1. What is the explanatory variable?

2. What is the response variable?

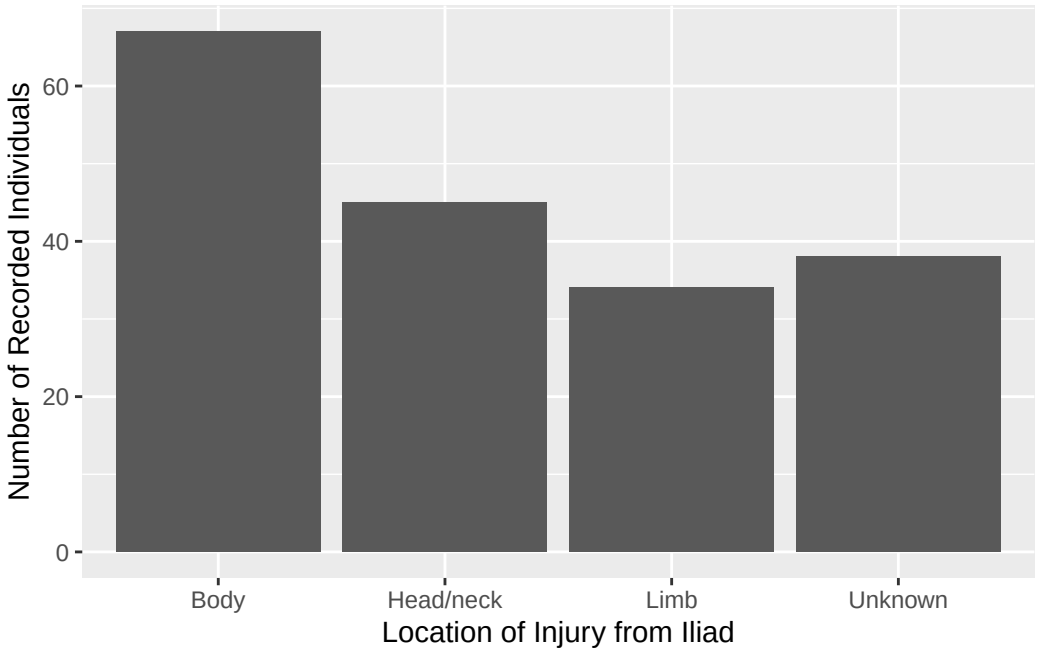
3. What is the scope of inference for this study?

Visualizing the Data

To visualize the relationship between **two** categorical variables, we need to add some color into our previous bar plots. Let's step through this process.

Last week, we started with a one variable bar plot that looked like this:

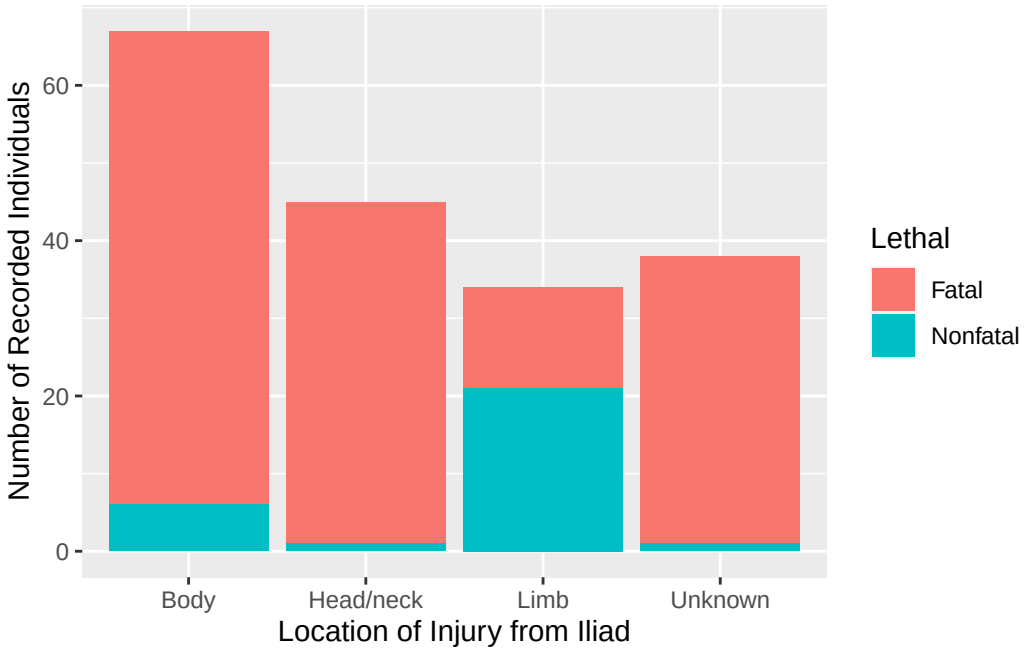
```
ggplot(data = iliad,
       mapping = aes(x = `Injury Site`)) +
  geom_bar() +
  labs(x = "Location of Injury from Iliad",
       y = "Number of Recorded Individuals")
```



The problem is, these bars don't show whether each injury included in the bar was fatal or nonfatal. To do this, we need to **fill** the bars with color, using the `fill` aesthetic.

In the code below, I've added a line that says `fill = Lethal`. This `fill`s each bar with two colors, one for "Fatal" injuries and one for "Nonfatal" injuries.

```r
ggplot(data = iliad,
       mapping = aes(x = `Injury Site`,
                     fill = Lethal)) +
  geom_bar() +
  labs(x = "Location of Injury from Iliad",
       y = "Number of Recorded Individuals")
```



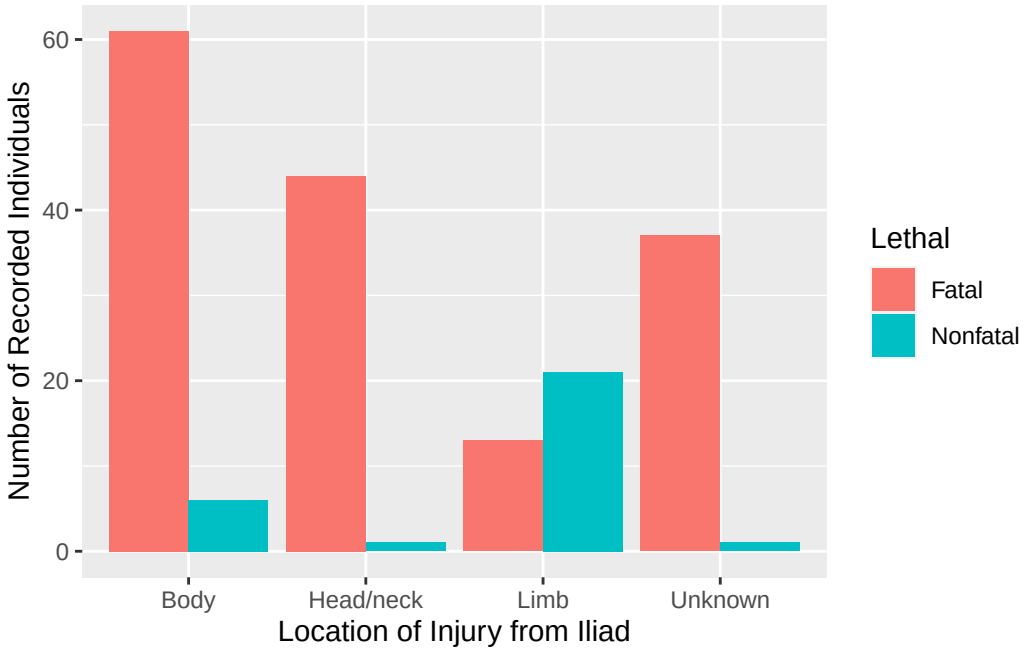
4. How would you describe the orientation of this bar plot? *Hint:* The three choices are filled, stacked, and dodged

If we want a different layout for our bar plot than what `geom_bar()` does by default, we will need to be more specific! This is where the `position` option in `geom_bar()` comes in handy. There are three options for how the bars can be `position`ed:

- `"stack"` (the default)
- `"dodge"` (to put the bars side-by-side)
- `"fill"` (to plot the proportions instead of frequencies)

Let's change the code to use a `dodge` position instead of a `stack` position.

```r
ggplot(data = iliad,
       mapping = aes(x = `Injury Site`,
                     fill = Lethal)) +
  geom_bar(position = "dodge") +
  labs(x = "Location of Injury from Iliad",
       y = "Number of Recorded Individuals")
```



5. Based on the plot does there appear to be an association between the variables? Explain your answer.

Summarizing the Data

Bar plots are an excellent tool for giving us a bigger picture of the relationship between two variables. However, sometimes we would like to see the **exact numbers** going into the bars. This is when it would be useful for us to have a table of the totals!

Like the bar plots, we will be adding to what we learned previously. In Lab 7, we made one variable tables using the following process:

```
myopia |>
  count(Light)
```

This gave us a table of how many observations (children) in the dataset slept with no light, a nightlight, or full light.

We could do something similar here:

```
iliad |>
  count(`Injury Site`)
```

Injury Site	n
Body	67
Head/neck	45
Limb	34
Unknown	38

But wait! This doesn't show how many of the 67 body injuries were fatal!

6. Take a guess at how we can modify the code above to give us counts of **both** the injury location and whether the injury was fatal.

```
iliad |>
  count(____________, ____________)
```

Nice work! You are correct, we need to put the **two** variables we are interested into the `count()` function. When we run the code you wrote this is what we get:

Injury Site	Lethal	n
Body	Fatal	61
Body	Nonfatal	6
Head/neck	Fatal	44
Head/neck	Nonfatal	1
Limb	Fatal	13
Limb	Nonfatal	21
Unknown	Fatal	37
Unknown	Nonfatal	1

7. Which injury location has the smallest number of observations?

8. Were there more fatal injuries or nonfatal injuries?

Unfortunately, these calculations are a bit difficult given the current layout of the table. What we need to do is pivot one of the variables (`Injury Site` or `Lethal`) to the columns instead of the rows. We can do this using a fancy tool called `pivot_wider()`!

I've added two lines of code to the previous table:

```
iliad |>
  count(`Injury Site`, Lethal) |>
  pivot_wider(names_from = `Injury Site`,
              values_from = n) |>
  adorn_totals(where = c("row", "col"))
```

Let me talk you through what each of these lines does.

```
pivot_wider(names_from = `Injury Site`,
            values_from = n) |>
```

takes the names from the **one** `Injury Site` column and makes **four** new columns based on the names from this column (`Body`, `Head/neck`, `Limb`, and `Unknown`). It fills each of those columns with the values found in the `n` column in the previous table.

```
adorn_totals(where = c("row", "col"))
```

takes the new table and adds a "Total" row at the bottom of the table and a "Total" column on the far right side.

The resulting table looks like this:

Lethal	Body	Head/neck	Limb	Unknown	Total
Fatal	61	44	13	37	155
Nonfatal	6	1	21	1	29
Total	67	45	34	38	184

Nice, right??? Use this new table to address the following questions.

9. What proportion of body injuries were fatal?

10. What proportion of limb injuries were fatal?

Chi-Squared Test of Independence

Similar to what we did last week with one categorical variable, we will be performing a Chi-Squared test to compare what we saw in the data to what we would have expected to see if the null hypothesis was true.

11. Write the null hypothesis for this study in words.

12. Using the research question, write the alternative hypothesis in words.

Expected Counts Under H_0

The Chi-Squared statistic (or X^2) has exactly the same formula to what we used last week. The only aspect that changes is how we get each cell's expected count.

To find the expected count for each cell in our two variable table, we use three pieces of information:

- the row total for that cell (denoted with an i)
- the column total for that cell (denoted with a j)
- the total sample size

We find the expected value of a cell using the following formula:

$$\frac{\text{row}_i \text{ total} \times \text{column}_j \text{ total}}{\text{total sample size}}$$

Once we have each of these, we can create a table of what frequencies we would have expected to see if H_0 was true.

13. I've gotten you started with the expected value table. Fill out the remainder of the table below!

Lethal	Body	Head/neck	Limb	Unknown
Fatal	$\frac{155*67}{184} = 56.44$			
Nonfatal		$\frac{29*45}{184} = 7.09$	$\frac{29*34}{184} = 5.36$	

Chi-Squared Statistic

Next, we compare each of our observed frequencies to what we would have expected if H_0 was true. We compare how far "off" our observed frequencies are from what was expected in a very specific way, using the following calculation:

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

14. Using the formula above, calculate how far "off" each of the cells in our observed table is from what was expected under the null hypothesis. *Hint:* You should be adding up **eight** numbers for your X^2 statistic!

15. Adding all of these differences together to obtain our observed X^2 statistic.

$X^2 =$

Sampling Distribution of X^2

In order for us to calculate our p-value—the probability of observing an X^2 statistic as or more extreme than what we got, if the null was true—we need a distribution of X^2 statistics that could have happened if H_0 was true.

Like all of our previous topics, there are two ways we can obtain this **sampling distribution**:

- using mathematical theory
- using computer simulation

Let's see how each of these works!

Theory-based Null Distribution

For the Chi-Squared test of independence we've been discussing, it can be mathematically shown that the distribution of X^2 statistics follows a χ^2 distribution with $(r-1) \times (c-1)$ degrees of freedom, where r is the number of rows and c is the number of columns in the table. Keep in mind the total row and column don't count!

16. How many degrees of freedom would be use for our χ^2 distribution?

The χ^2 Distribution

We've seen the χ^2 distribution before in the context of the goodness-of-fit test. That is the same distribution we are using here! Recall, the χ^2 distribution has only positive values and the shape of the distribution is controlled by its degrees of freedom.

Conditions for Using a χ^2 Distribution

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

If the condition about expected cell counts is violated, we are forced to use a simulation-based method.

17. Are the conditions for using a χ^2 distribution to approximate the sampling distribution violated?

Using a χ^2 Distribution to Find the p-value

If we decided in #17 that it is not unreasonable for us to use the χ^2 distribution, then we can use `R` to find our p-value.

We will use the `pchisq()` function (which you saw last week). This function has three inputs:

- the observed X^2 statistic
- how many degrees of freedom should be used for the χ^2 distribution
- if the lower tail (left tail) should be used when calculating the p-value

18. Using the values you calculated before, fill in the code below:

```
pchisq(_________, df = _____, lower.tail = FALSE)
```

Running the code you just wrote in `R` gave me a p-value of approximately 0.

19. Based on this p-value what conclusion would you reach regarding the null hypothesis? *Hint:* Go back and see what you wrote for your null and alternative hypotheses in #11 and #12!

Simulated / Permuted Null Distribution

If the condition for expected counts is violated, we cannot use a χ^2 distribution to approximate what the true sampling distribution looks like. Instead, we need to use computer simulation to obtain our p-value.

Like all of the previous times, we can think about what the computer is doing using cards. Keep in mind, we are **assuming the null hypothesis is true**, which suggests there is no relationship between the location of someone's injury and whether they lived or died.

To carry out **one** simulation we need to do the following steps:

Step 1: Write the response (died / didn't die) and the location of the injury (body / head / limb / unknown) on _____ cards.

Step 2: Assume the null hypothesis is true and:

Step 3: Create a new dataset that could have happened if H_0 was true by:

Step 4: Create a new table of counts for the shuffled cards.

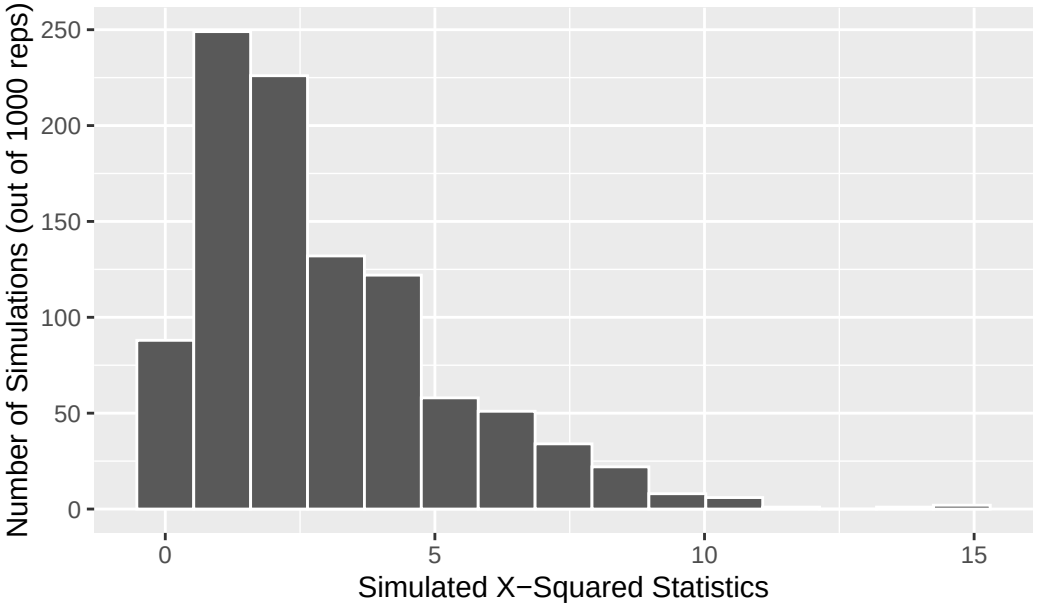
Keep in mind the **column totals** and the **row totals** will stay the **same** (there were only 34 people with limb injuries, and only 155 people who died). However, the **cell values** will **change** (we won't always have 61 fatalities for body injuries).

Step 5: Calculate the X^2 statistic for the simulation.

Because the row and column totals stay the same, the table of expected counts will be the same for **every** simulation! That's handy!

Step 6: Plot the simulated X^2 statistic on the distribution

Alright, after carrying out this process, I obtained the following distribution.



20. Draw a line where the observed X^2 statistic falls on this distribution.

21. Estimate the p-value for testing if there is a relationship between the location of an injury and whether the individual survived.

22. Based on your p-value, what would you conclude for your null and alternative hypotheses? (Look back at what you said for #11 and #12)!

23. Did you reach similar conclusions using theory-based methods? Why do you believe that is the case?