Freya Huang (AE)

Tapeworm

(1) Does the tapeworm dataset represent an observational or experimental dataset? Which variable(s) are predictor variables? Which are response variables? Explain your answers in a total of 2-4 sentences (5 points).

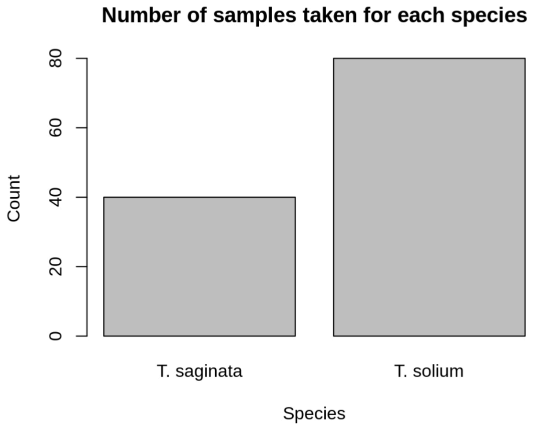
Tapeworm dataset represents an experimental dataset. This graph includes data that determine tapeworm's growth under different temperature conditions, in the presence of abundant food. In this case, different temperatures are predictor variables since temperatures make the effects, and growth is response variables since it receives effects.

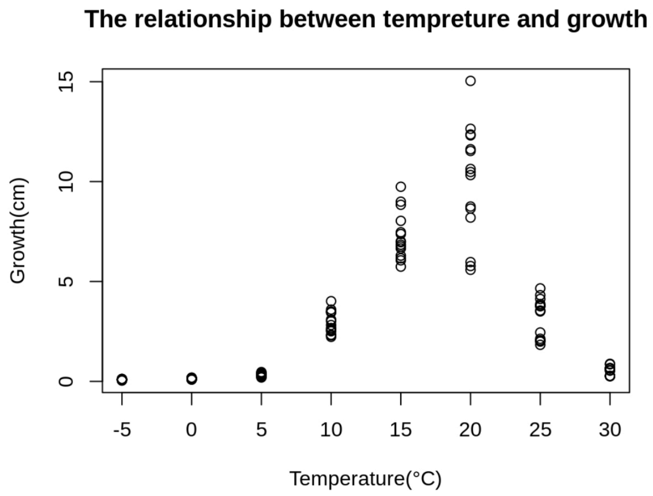
(2) `View` the tapeworm dataset. Which columns are categorical? discrete? continuous? binary? Explain in 1-2 sentences why `temperature` is difficult to categorize. (5 points)

Species is categorical, but growth, temperature, and trail are numerical columns. Among numerical datasets, the trail is discrete, but temperature and growth are continuous. Temperature is difficult to categorize because in the dataset, the data presented in the "temperature" column are all whole numbers, which may mislead people to understand those as discrete variables. However, temperature can be count as 5.1°C, it can be continuous, so it is a continuous column.

(3) The researchers did not balance their samples with respect to species. Make a bar chart showing the number of samples taken for each species. Interpret the bar chart in 1-2 sentences.

The bar chart illustrates the number of samples taken for each species. The x-axis includes two columns, or species: T. saginata and T. solium. Y-axis counts how many tapeworm samples are taken from different species. T. saginata is counted to take 40 samples, and T. Solium is counted to take 80 samples.





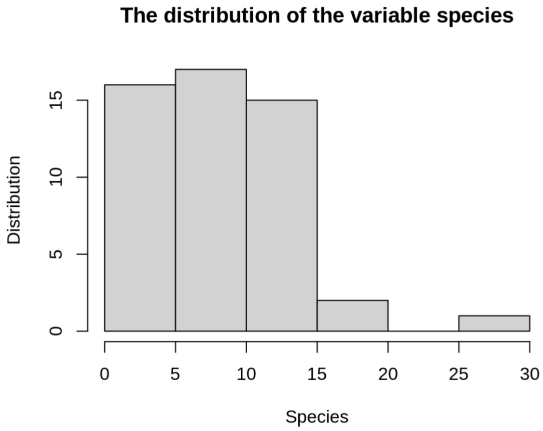
(4) Make a scatterplot showing the relationship between `temperature` and `growth`. Interpret the relationship between `temperature` and `growth` in 1-2 sentences (5 points).

According to the scatter plot, we see that after 10 days, tapeworms that lived in temperatures equal or less than 5°C grew to be very small, most of them are less than 1cm. While tapeworms live in temperatures between 10°C to 25°C, they have grew bigger significantly: many of them are bigger than 5cm. Especially at the temperature of 20°C, the length of one tapeworm is 15cm. However, if tapeworms live in a temperature hotter than 25°C, their lengths drop to less than 5cm again. This plot illustrates the best temperature for tapeworms to grow is 20°C.

Invasive Species

(5) `View` the invasive species dataset. Calculate the mean, standard deviation, and max of the number of invasive species detected. Interpret the summary statistics in 1-2 sentences (5 points)

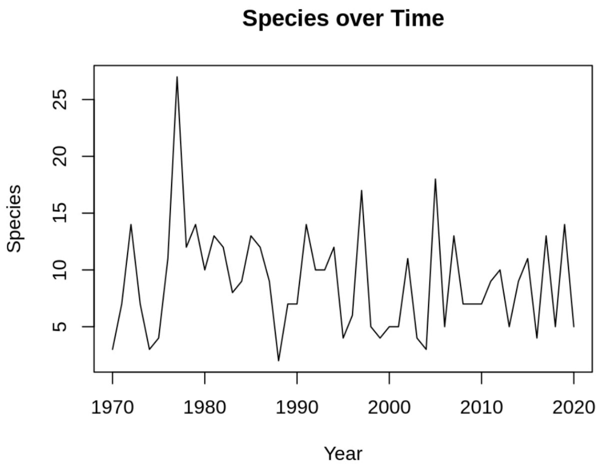
The mean of the number of invasive species detected is 8.941 species; The standard deviation of the number of invasive species detected is 4.726 species; and the Max of the number of invasive species detected is 27 species.



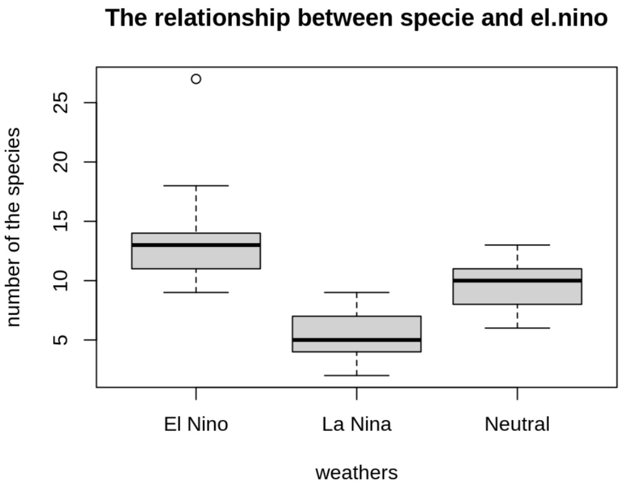
(6) Make a histogram showing the distribution of the variable `species` from the invasive species dataset. Interpret the distribution in 1-2 sentences (5 points).

This histogram is a left-skewed diagram. Most of the species are distributed on the right side. It means that most of the distributions happen between species of 15 and smaller than 15. After species of 15, there were fever distributions happened.

(7) Make a line plot showing `species` over time. Interpret the changes in number of newly detected invasive species over time in 1-2 sentences (5 points).
In general, the number of new invasive species detected in the U.S shows a decreasing tendency while the number of detected invasive species has a large fluctuation over the years. Besides, the max is between 1970s to 1980s, and the min is between 1980s to 1990s.



(8) Make a box plot showing the relationship between `species` and `el.nino`. Interpret the relationship in 1-2 sentences (5 points).
In this box plot, under the weather of El Niño, the number of invasive species is higher since the median of this box is higher than the other two. Under the weather of La Niña, the number of invasive species is generally lower than the number of invasive species under the neutral weather.



Independent Learning Problem:
Read the attached document about sampling bias and randomized controlled trials.

(1) Give an example of a study where self-selection bias might occur and explain why in 1-3 sentences (5 points).
An example can demonstrate the self-selection bias. Researchers hold a study focuses on how to educate children, and they asked different people to participate in their study. People who want to be good parents in the future are more likely to participate, while others who don't care about children's development will not participate. This is a typical example because, in the study, participants can choose whether they want to take a part in the study, which affected the result to be not equivalent, and in turn, not balanced.

(2) Give an example of a study where exclusion bias might occur and explain why in 1-3 sentences (5 points).
Researchers want to count how many people in an undeveloped town can read, so they write down the detailed information in an email with the poll attached below, and send it to people living in this town. It turns out that everyone can read. However, it is not an efficient result since the exclusion bias occurred. Only people who read the emails, which means they can read, know how to vote in polls. People who cannot read, cannot vote in the poll, either. In this case, certain individuals are unable to participate in the study for reasons other than their wishes.

(3) In class we emphasized the idea of the population that your sample represents. Explain what a population is and how each type of bias can influence the representativeness of your sample (5 points).

A population represents a collection that contains all individuals in a group. It is the opposite of a sample. Both biases would cause the results to be ineffective.

In the example for question 1, the self-selection bias occurs because there may only have parents who care about their children would participate in the study, leading to a sample that doesn't contain parents who don't care about children's development. However, in the population, there are other parents like those. So, this sample is not a good representative of the population. In the example for question 2, the exclusion bias occurs because there is defiantly a portion of people who cannot read in an undeveloped town. However, it doesn't show up in the result. In this way, this sample is not a good representative of the population.