Week 4 Exercises: Statistics for Data Science

Part 1: Penguins

For this exercise, we are going to revisit the penguins dataset contained in penguins.csv.

Read this file into a dataset named `penguins` and then filter down to two datasets, `adelie` and `chinstrap` using the following code:

```
adelie = penguins[penguins['species'] == 'Adelie'].dropna()
chinstrap = penguins[penguins['species'] == 'Chinstrap'].dropna()
```

1. Find the average `flipper_length_mm` by species (using the full dataset).
2. Build a 95% confidence interval for the mean `flipper_length_mm` for adelie penguins.
3. Build a 95% confidence interval for the mean `flipper_length_mm` for chinstrap penguins.
4. Based on the confidence intervals you build, do you feel like it is safe to conclude that the population of Chinstrap penguins has a higher average flipper length than the population average flipper length for Adelie penguins?
5. Find the average `body_mass_g` by species (using the full dataset).
6. Build a 95% confidence interval for the mean `body_mass_g` for adelie penguins.
7. Build a 95% confidence interval for the mean `body_mass_g` for chinstrap penguins.
8. Based on the confidence intervals you build, do you feel like it is safe to conclude that the population of Chinstrap penguins has a higher average body mass than the population average body mass for Adelie penguins?

Part 2: Conceptual

You are planning to build a confidence interval for the mean of some numeric variable. You have not yet gathered your sample, but based on past history, you expect that your sample will have a standard deviation around 10.

1. If you are going to build a 95% confidence interval and want to end up with a margin of error less than 10, approximately what sized sample do you need?
2. If you are going to build a 95% confidence interval and want to end up with a margin of error less than 1, approximately what sized sample do you need?
3. If you are going to build a 95% confidence interval and want to end up with a margin of error less than 0.1, approximately what sized sample do you need?
4. What do you notice about the relationship between the desired margin of error and the needed sample size?

5. How do your answers change if you expect the standard deviation to be around 5 instead of 10?

Part 3: Squirrels

Read the squirrel census data (2018_Central_Park_Squirrel_Census_-_Squirrel_Data.csv) into a DataFrame named `squirrels`.

1. For what proportion of encounters was it the case that the squirrel was seen approaching (as indicated in the `Approaches` column)?
2. Build a 95% confidence interval for the proportion of squirrel encounters where the squirrel approaches. What is the margin of error of this confidence interval?
3. Build a 95% confidence interval for the proportion of squirrel encounters where the squirrel is seen running (as indicated in the `Running` column). What is the margin of error of this confidence interval? How does the margin of error in this case differ from the margin of error in the previous question? Why does this happen?

For the next set of questions, divide the data into encounters into morning encounters and afternoon encounters using the following code:

```
am_squirrels = squirrels[squirrels['Shift'] == 'AM']
pm_squirrels = squirrels[squirrels['Shift'] == 'PM']
```

4. First, let's compare the morning and evening encounters with respect to the `Foraging` column. For what proportion of morning encounters was the squirrel seen foraging? For what proportion of afternoon encounters was the squirrel seen foraging?
5. Build two confidence intervals, one for the morning encounters and one for the afternoon encounters, for the proportion of encounters where the squirrel was seen foraging. Based on these confidence intervals, do you think it is safe to conclude that for encounters in the afternoon, the squirrel is more likely to be foraging compared to encounters in the morning?
6. Repeat the previous question, but for the `Runs from` column.