Problem Set 2

This problem set is due on February 19, 2026. Upload your write up and R files to Canvas by 5pm on the due date. You may work together, but you must turn in separately written (unique) write ups and code. All answers need to be written in clear and concise sentences.

Make sure you have the load any packages you need at the start of your report

```
library(readr) ## for reading data
library(dplyr) ## for working with data using reframe and mutate
library(knitr) ## for making nice looking tables
library(ggplot2) ## for making nice looking plots
```

1. Heading into the 2024 election, RealClear Politics (RCP) published several polls for the state of Michigan. Results of their listed polls that started on or after October 27th are listed below

Harris	Trump
48	50
47	47
47	47
50	48
47	47
50	48
49	48
51	48
48	48

   (a) (5 points) Create a data frame in R using these data. Are these paired or unpaired data? Explain how you know and create a new column that is Trump support minus Harris support.

   (b) (5 points) Suppose that each row represents an independent draw from the sampling distribution of $\hat{\mu}_{\text{diff}}$. Describe the true mean and variance of this sampling distribution. Is it normal? Bernoulli? Something else?

(c) (10 points) The final vote margin in the election was Trump +1.4. Using your answer to parts a and b, find the probability that the true vote difference was Trump at least +1.4. Are you surprised? Why or why not?

2. In this problem, we will consider data on domestic terrorism around the world. The data are in the file `terrorism1.csv`, load these data in R. Your client wants to know if democracies experience more or less terrorism than autocracies and whether terrorist attacks within democracies are more or less deadly.

   (a) (5 points) Find out what the unit of observation is in this data set by looking at the data. Use any commands you know to determine the range of years in the data.

   (b) (10 points) Produce a table of summary statistics (e.g., minimum, mean, standard deviation, maximum) and appropriate plots to describe the variables `domAttacks` and `nkill`. This former counts the number of observed terrorism attacks observed in that observation, the latter counts the number of deaths due to domestic terrorism in that observation.

   (c) (5 points) Create a new variable called `deaths.per.attack` equal to the number of deaths divided by the number of attacks.

   (d) (5 points) We will assume that each observation is independent, how believable do you think that is? What are the consequences of non-independent observations?

   (e) (5 points) The first question your client has is: Do democracies experience more or fewer attacks, on average, than autocracies? Write out the null and alternative hypotheses for this question. Are the data paired? Is a z test appropriate here?

   (f) (10 points) Note that the variable `demo` is binary: a 1 denotes a democracy and 0 an autocracy. Create a visualization to compare `domAttacks` for democracies and autocracies. What differences do you see?

   (g) (10 points) Test hypothesis in part (e). What do you find? Use a significance level of your own choosing.

   (h) (10 points) Repeat parts (e-g) but with `deaths.per.attack` instead of `domAttacks`. What do you find?

(i) (10 points) In a **short** paragraph answer the clients two questions and discuss your findings. What is interesting (or not) about these findings? Can we draw a causal conclusion about democracy and terrorism? Why or why not?

3. (10 points) You are given the following hypotheses:

$$H_0 : \mu = 55$$
$$H_A : \mu \neq 55$$

We know that the sample variance is 16 and the sample size is 100. For what two possible sample means would the p-value be equal to 0.05? Assume an independent sample from an approximately normal population. (HINT: Remember that function `qnorm` does the opposite of `pnorm`. It takes p values or percentiles and returns z values.)