DSA 8010 - Project 3 (10 points)

Please submit your project through Canvas by 11:59 pm on Tuesday, Nov 19. You may work in a group of size 2-4, but are not required to. If you work in a group, each person in the group should submit the report in Canvas. Include all group members' names at the top.

Instructions

You are to perform two analyses, worth 5 points each. Choose two questions from the three described below. Each question gives a data set and asks an open-ended "key question" about what the researcher hopes to learn from the data. Your job is to write a short analysis of the data that addresses this key question. Turn in your answer in a pdf, .docx, or similar format.

For each question, your answer should include each of the following elements:

1. A brief **descriptive summary** of the data that addresses the key question. Include summary statistics and an appropriate plot or table. For full credit, make sure your calculations are integrated into the text. Give your description in paragraph form.

2. An appropriate **inferential analysis** to address the key question. This may be a confidence interval, hypothesis test, or both. You may use the α value/confidence level of your choosing. If you perform a hypothesis test, state the hypotheses, test statistic, p-value, decision, and conclusion. If you create a confidence interval, state the confidence level and provide an interpretation of the interval.

3. A thoughtful **conclusion** based on the information presented in items 1-3. The conclusion should summarize your findings with respect to the key question. You should also discuss any limitations or complications of your analysis. This must include at least one of the following:

    - an assessment of whether the modeling assumptions are reasonable. This may include making histograms or normal quantile plots or checking the numbers of successes/failures.
    - discussion of whether the data are representative of a well-defined population. Consider, for example, any information about how the data were collected.
    - discussion of any unusual features of the data. For example, do outliers affect your results?

Most answers will contain 2-4 paragraphs and 1-2 plots/tables per question. Try to keep your answers concise. Make sure to answer all questions.

Q1: e-cigarettes

A researcher is studying the effect of e-cigarettes as an aid to smoking cessation. In her study, $n = 29$ participants with an intention to quiz smoking were given free e-cigarettes and e-liquids for a total of six weeks. They reported their cigarettes smoked per day (CPD) at the beginning of the trial and on a weekly basis for the duration of the study. Some additional characteristics of the participants and their behaviors were also reported.

The file `ecigarettes.csv` gives the cigarettes per day at baseline (CPD_BL) and week 6 (CPD_W6) for each study participant.

Key question: Do the data suggest that the participants' cigarettes per day decreased over the course of the study?

Use statistical methods to address the key question. **In your conclusion, include a brief discussion of this question in addition to other required components: based on the information provided, does this study design allow for effective evaluation of whether e-cigarette usage leads to decreased cigarettes per day? Why or why not?

Q2: heifer feed

A team of animal science researchers want to investigate the effects of different supplemental feeding methods on young heifers. In this study, all cows in an experimental farm were randomly assigned to be given a feed supplement either from a shared trough with other cows (`Treatment=Group`) or from an automatic precision feeder that dispensed a pre-determined amount to each heifer as it approached the feeder (`Treatment=Precision`). The supplements were given at two levels, 0.5% and 1% of the cows' bodyweights.

The cows' weights (lbs) were measured once a month for about 6 months. The weights and identifiers for each cow are found in the data set `heifer_mod.csv`. The researchers are interested in seeing how the total weight gain (`tgain`) differs across the two types of feeding methods.

Key question: On average, did cows using Group feeding have a different total weight gain than cows using the Precision feeding in the low feed level (`level=0.5`)? What about in the high feed level (`level=1`)? (Perform the descriptive and inferential analyses separately for the two levels.)

Q3: teen dating violence

A study of teenagers in rural South Carolina surveyed high-schoolers to ask about whether they had experienced different types of violence in dating relationships. This was a longitudinal study in which the same students were surveyed four times (in grades 9, 10, 11, 12) and each time asked the same questions about their experiences in dating relationships. For some students, the survey team was not able to obtain survey response for all four consecutive years.

For the purpose of this project, consider a student to be "lost to follow up" if their response to the grade 12 survey is missing. Certain groups of students might be more likely to be lost to follow up, and so there is a possibility that the survey responses at grade 12 are subject to nonresponse bias because these groups are underrepresented.

The file `dating_survey2.csv` contains information on gender, race/ethnicity, and maternal education level for 580 students as well as their responses to the question "In the past year, were you the victim of physical violence in a dating relationship?" in grades 9, 10, 11, and 12. Missing responses are shown as NA in the data set. Maternal education was recorded using the following values for highest educational attainment of student's mother: 1=less than high school diploma; 2=high school diploma or post-secondary education.

Key question: Were students of one maternal education level more likely to be lost to follow-up than those from another?

**You will need to pre-process the data before the analysis by finding out which students are lost to follow up.

Rubric for all questions

Category	Points
The results are presented clearly in a readable, narrative form. The writing is clear and concise. All quantitative results, including plots, tables, and summary statistics are referenced and interpreted in the text. Captions, graph titles, and axis labels are included where appropriate.	1
A brief and informative descriptive summary is given. The summary statistics are correct and their relevance to the key question is described in the text.	1
An appropriate inferential procedure is chosen and implemented correctly.	2
A thoughtful and accurate verbal conclusion is given. It summarizes findings with respect to the key question. Limitations or complications of analysis are discussed, using at least one of the three points given above.	1