Midterm 2

Instructions:

Download the diabetes.csv data file. You can open an Rmd on your desktop RStudio application or using posit.cloud.

Answer all questions to the best of your ability. ***Do not forget to self-assess your performance after you completed the midterm!***

The Pima people of Arizona initially supported themselves through farming, but around 1900, water diversion by white settlers led to a collapse of their agricultural way of life. This led to food scarcity, reduced physical labor, and a high-fat diet. This lifestyle shift coincided with a dramatic rise in diabetes among the Pimas, leading to exceptionally high diabetes prevalence, with rates continuing to increase and affecting nearly half of Pimas over age 35 by the 1970s. Epidemiological evidence indicates that Type 2 Diabetes results from interaction of genetic and environmental factors, some of which you will investigate during this exam.

The Pima Indian Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information of 500 randomly sampled women from a population near Phoenix, Arizona, USA.

Variable	**Description**
pregnancies	Number of times pregnant
glucose	Oral glucose test result *(values during the test may range between 90-200*
blood_pressure	Diastolic blood pressure (mmHg) *(average BP is around 80)*
insulin	2-hour serum insulin (mm U/ml) *(in extreme cases can range from 0 to over 300.*
bmi	Body mass index *(normal BMI range for females is between 18.5-24.9)*
age	Age in years
outcome	1 = diabetes, 0 = no diabetes

Section 1: Study Design

1. How does the use of a specific subpopulation (Pima Indian women aged 21 and older) impact the generalizability of the study's findings on diabetes risk factors to a larger population?
2. Selection Bias: What type of women from the study are most likely to choose to be a part of the study and how does this impact the generalizability of the results?

Section 2: Exploratory Data Analysis (EDA)

3. Use the `summary()` function to get a summary of all of the variables. Comment on any unusual values in the data. How might these affect your analysis?
4. Decide how you will handle the unusual values. For example, will you remove them and if so, how? Justify your decision in a few sentences.

However you have decided to handle the missing values, do so in your code here.

Data cleaning: make sure any categorical variables are coerced to factors before moving on to the next questions.

5. Create a publication-quality table (e.g., using `tbl_summary`, `kable`, or some other function) of descriptive statistics for each variable in the dataset, stratified by outcome. Quantitative variables should be summarized with the mean (sd) and categorical variables should be summarized with n (%).
6. In a few sentences, comment on any patterns you notice in the table between those diagnosed with diabetes and those not diagnosed.
7. Create a well-labeled plot to visualize the relationship between BMI and diabetes status (outcome).
8. Describe the plot in words, commenting on any notable features. You may consider including some summary statistics to add more information in your description. Imagine you are describing the plot to someone who cannot see it.

Section 3: Data Analysis

9. Perform a 5-step hypothesis test to determine whether there is evidence of a difference in the proportion of diabetes diagnoses between women under 30 and women aged 30 or older. You will need to create a new variable that categorizes age into under 30 and 30+. **Note: write your hypotheses in symbols.*
10. Test whether there is a significant difference in the average BMI between those with diabetes and those without diabetes using a confidence interval at the 1% significance level. Write your hypotheses (in symbols*), check the necessary conditions, calculate the confidence interval, and write a conclusion in the context of the problem.

11. Finally, test whether there is an association between the number of pregnancies and diabetes status (outcome). Perform all 5 steps.

Section 4: Conclusion

12. Based on the tests and analyses you have conducted, summarize your conclusions regarding the relationships between diabetes and age, BMI, and number of pregnancies.
13. How might these findings inform healthcare policy or targeted interventions to reduce the rate of diabetes among the Pima Indian (female) population?

Reflection Questions

14. Was there anything you found difficult with this exam? What topics (if any) do you feel you still need more work on?
15. Give yourself a rating for this exam using the EMRN rubric.

E - Excellent; M - Meeting expectations; R - Revision needed; N - Not accessible

*LaTeX code for mathematical symbols if you would like to use this

`$H_0$` = H_0 `$H_A$` = H_A

`$\bar{x}$` = $\bar{x}$

`$\mu$` = μ

`$\mu_1$` = μ_1

`$p_1$` = p_1

`$\hat{p}$` = $\hat{p}$