COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

In-Class Assignment 9

Using the sqldf Package in R and PROC SQL in SAS

Read the scenarios below, then follow the instructions and answer the questions that follow. Submit your solutions in the **NHANES.R** file and the **In-Class Assignment 9.sas** to Canvas by the deadline listed above. Save your file frequently to avoid losing work!

Part 1: R

Scenario:

You will be analyzing data from NHANES using R software. Download the NHANES csv files and write SQL queries using the sqldf package to help answer some of the analytic questions.

Queries:

1. Write one SQL query that would allow you to fill out the following table. Assign the query results to an object called **table1.** Utilize the ROUND( ) function to round mean age to 1 decimal place (example syntax: ROUND(variable, 1)) and use the column aliases **freq** and **mean_age**.

Table 1

Race	Frequency	Mean Age
Mexican American		
Other Hispanic		
Non-Hispanic White		
Non-Hispanic Black		
Non-Hispanic Asian		
Other		

2. Now show the distribution of race by gender and display the race/gender combinations from highest to lowest frequency. Use the column alias **freq**. Note: when using SQL in R, you can refer to column aliases outside of the SELECT clause.

3. Count the number of women who were pregnant at the time of screening. Use the column alias **preg_at_screen**.

4. How many men refused to provide annual household income? Use the column alias **num_refused**.

In-Class Assignment 9

5. What is the mean LDL level (mg/dL) for men and women? Use column alias **mean_ldl** and round results to 1 decimal place. Also use table aliases.

6. Display the minimum and maximum triglyceride levels (mmol/L) for each race. Use column aliases min_tri and max_tri.

7. Create a new data frame that can be used for future analyses that combines all demographic data and any matching triglyceride data. Call it demo_tri.

Part 2: SAS

Scenario:

You will be analyzing data about populations in different countries using SAS software. Download the SAS data and editor files and write the SQL queries below.

Queries:

1. Write a SQL query to display a list of the distinct continents represented in the data.

2. List the total population of each continent, ordered from highest to lowest.

3. In the absence of a data dictionary, write a query that would help you determine which continent is represented by the continent number that had the highest population in query 2.

4. Write a query to calculate the number of individuals who reside in urban areas in each country and save it into a new table called **sql.urban**. Name the variable you calculate as **total_urban_pop**.

5. Write a query to display the maximum population value in the dataset.

6. Use the query you wrote in #5 as a subquery to find the country or countries linked to the highest population value.

Guidelines:

- Follow style guidelines for capitalizing keywords (and optionally for line breaks and indents)

- When more than one table is involved in a query, use table aliases