**Biostatistics 203A**

**Final Project**

Due: December 15, 2023

The final project will present you with an opportunity to use some of the skills you have learned throughout the quarter in Biostatistics 203A including manipulating and summarizing data, creating graphs and tables, conducting statistical tests, simulating data, and composing a report.

In the first part of this final project we will be analyzing data from a multi-wave survey called How Couples Meet and Stay Together (HCMST). The intent of the survey was to learn more about how Americans met their spouses and romantic partners and to compare traditional and non-traditional couples. During the first wave, couples were asked about their relationship status and the sample we will be analyzing includes only those respondents that indicated having a spouse or other romantic partnership at this first wave. Data for this study were obtained from ICPSR (Inter-University Consortium for Political and Social Research), an excellent source for interesting publicly-available data sets. If you are interested, more information can be found at the ICPSR website and here is the HCMST citation:

Rosenfeld, Michael J., Thomas, Reuben J., and Falcon, Maja. How Couples Meet and Stay Together (HCMST), Wave 1 2009, Wave 2 2010, Wave 3 2011, Wave 4 2013, Wave 5 2015, United States. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-03-18. <https://doi.org/10.3886/ICPSR30103.v8>

A codebook has been uploaded to the course website along with a master data set (**HCMST.sas7bdat**), a **CaseSubset.csv** file, and a **Final Project SubsetNumber Assignment.csv** file. You will need to use this codebook in order to understand the variables and data you will be using in Step 1 below. The master data set you receive (**HCMST.sas7bdat**) represents a subset of both the observations and the variables from the original downloaded data.

As with the mid-quarter project, you will need to use the **Final Project SubsetNumber Assignment.csv** file to first find your unique SubsetNumber. You will then need to use the **CaseSubset.csv** file to obtain the subset of CASEID\_NEW (the unique identifier for respondents in the HCMST data) values that you will be working with for Step 1. A good first step is to filter/subset the **HCMST.sas7bdat** file based on your subset of CASEID\_NEW values and to work only with this subsetted data going forward.

In the second part of this final project, you will be asked to create an R function and use that function to explore the implications of conducting multiple hypothesis tests.

The project can be broken down into 3 Steps:

1. Write a SAS macro to examine the association of a single categorical variable taking values from two categories with each of the following variables:

* Where did you meet your partner? [Q31\_1 – Q31\_9]
* Who introduced you to your partner? [Q33\_1 – Q33\_7]
* An indicator of whether or not respondent age and partner age are ≤ 5 years apart. [PPAGE and Q9]

Among the set of variables listed above, please calculate chi-square tests to examine associations for the categorical variables.

The SAS macro should take exactly 4 arguments (in this order) corresponding to

1. The SAS variable name for the categorical variable (for instance, GENDER)
2. A label that corresponds to the name which will be used to ease interpretation of output tables and plots (for instance, “Gender”)
3. A user-defined format that should be applied to the categorical variable to ensure that formatted values appear in the output tables rather than underlying values (for instance, “Female” instead of 2). Note: This means that you will want to define the format outside of the macro and read it in before executing the macro.
4. The name of the SAS data set that contains the categorical variable and all the variables listed above

When executed, your macro should return a single table containing counts and percentages you feel might be helpful along with p-values from all the chi-square tests conducted. You may want to test your macro using variables such as PPGENDER or MARRIED.

1. Create a new variable that serves as an indicator of whether the respondent chose “Excellent” in response to the relationship quality question. This information can be found within the RELATIONSHIP\_QUALITY variable. Let the response of Excellent receive a value of 1 and any other non-missing response receive a value of 0 for this new indicator variable. Provide this new variable (along with appropriate name and format arguments) as arguments to the macro you created in Step 1. Based on the results in this table, provide your answer to each of the following questions in the document you submit:
2. Of all the meeting locations, which one corresponded to the highest percentage of respondents endorsing Excellent relationship quality?
3. Of all the different routes of introduction, which one corresponded to the highest percentage of respondents endorsing Excellent relationship quality?
4. What percentage of respondents whose partners were a similar age endorsed Excellent relationship quality? Did this percentage differ significantly (p < 0.05) relative to respondents whose partners were not a similar age?

For the response to (I.) above, use R to create a proportional (100%) stacked bar graph comparing the distribution of responses to the original RELATIONSHIP\_QUALITY item among those who met at the stated location relative to those who did not meet at the stated location. The original RELATIONSHIP\_QUALITY variable contains 5 different response options. Please ensure labels, axes, and titles are appropriate and useful. Also, please include a well-formatted legend. Place your graph in the document you submit.

1. Create an R function that will allow us to simulate the conduct of multiple chi-square tests (as conducted in Steps 1 and 2) under various assumptions. When executed, the R function you create should complete the following:
   1. Simulate *N* = *n*1 + *n*2 observations
      1. *n1* from a binomial distribution with probability of success *p*1
      2. *n2* from a binomial distribution with probability of success *p*2
   2. Repeat part (a) *M* times, such that you have M sets of results that could each be displayed within a2 x 2 contingency table. Using these 2 x 2 contingency tables, conduct *M* chi-square tests and save the corresponding *p*-values to an R object.
   3. Determine how many of the *M* *p*-values indicate a statistically significant association at the *p* < 0.05 level.

The function you create should perform the Steps (a)-(c) above *G* times and should output a single vector of length *G*. Each element in the output vector should contain the number of statistically significant test results (out of *M* possible). The function you create should be flexible in that it should take the following arguments:

* *n1* (the number of observations in group 1)
* *n2* (the number of observations in group 2)
* *p*1 (probability of success in group 1)
* *p*2 (probability of success in group 2)
* *M* (the number of chi-square tests)
* *G* (the number of replications)

Run the R function with the following inputs:

* *n1* = *n2* = 100
* *p1* = *p2* = 0.20
* *M* = 20
* *G* = 1000

Based on the output vector obtained after running the function above, determine the proportion of the *G* replications that resulted in at least one statistically significant result. Then, repeat the process above with the same inputs except for M, which you will have range from 1 to 30 (one iteration per integer 1, 2, . . 30). Save the proportions you obtain from each iteration of the process and create a line graph with *M* on the x-axis and the resulting proportion on the y-axis. In 2-3 sentences, describe the results you obtained and how they may have implications for situations in which multiple hypothesis tests are being conducted.

What you will submit:

1. A word document containing:
   1. Written responses to items I, II, III in Step 2
   2. The proportional (100%) stacked bar graph generated in Step 2
   3. The line graph generated in Step 3
   4. The 2-3 sentences you wrote as part of Step 3
2. A SAS syntax file containing the SAS macro described in Step 1.
3. An R Script containing the R function described in Step 3.

**All project components are to be uploaded to BruinLearn no later than 5:00 PM on Friday, December 15th**.