Biostatistics 203A
Final Project
Due: December 13, 2019

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 brief written 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 Steps 1 and 2 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 Steps 1-2. 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 4 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 and partner indicated the same race. [RESPONDENT_RACE and PARTNER_RACE] (researchers refer to this as assortive mating)
- An indicator of whether or not respondent age and partner age are ≤ 5 years apart. [PPAGE and Q9]
- Relationship quality [RELATIONSHIP_QUALITY]

Among the set of variables listed above, please calculate chi-square tests to examine associations for the categorical variables. Relationship quality is an ordered categorical variable and you will want to present the results from a chi-square test for trend.

The SAS macro should take at least 4 arguments 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

You may find it helpful to include up to 4 additional arguments corresponding to the values that the categorical variable takes (for instance, 0 and 1) and the labels that correspond to those values (for instance, "Male" and "Female"). Please include no more than 8 arguments total. When executed, your macro should return a single table or a small number of tables 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.

2. Create a new variable that indicates whether an individual remained with his/her partner through Wave 4 of data collection or if the couple broke up any time prior to Wave 4. To do this, you will need to use the W2_BROKE_UP, W3_BROKE_UP, and W4_BROKE_UP variables. A break up having been indicated at any of these three waves will be considered sufficient to classify an individual as having broken up with his/her partner. To be classified in the "stayed together" group, however, an individual must have a non-missing record at wave 4 that indicates no breakup (and no death). Once you have created this new variable, create a new data set that contains only records that have a non-missing value for this new variable. Provide this new data set and new variable (along with appropriate name and format arguments) as arguments to the macro you created in Step 1. Place the table or tables that are output in a word document following a very brief description of the results. You do not need to discuss the methods, simply provide a brief summary of the pertinent findings.

3. Create an R function that will allow us to simulate the conduct of multiple chi-square tests (as done in Step 2) under various assumptions. When executed, the R function you create should complete the following:
    a. Simulate $N = n_1 + n_2$ observations

i. n_1 from a binomial distribution with probability of success p_1
ii. n_2 from a binomial distribution with probability of success p_1

b. Repeat part (a) *M* times, such that you have M sets of results that could each be displayed within a 2 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.
c. 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:

- n_1 (the number of observations in group 1)
- n_2 (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:

- $n_1 = n_2 = 100$
- $p_1 = p_2 = 0.20$
- $M = 20$
- $G = 1000$

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 plot 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.

4. In this last Step, you will be using the function you created in Step 3 to mimic the results we obtained when analyzing the HCMST data in Step 2. Execute the R function you created by providing the following inputs:

- n_1 = number of individuals in your "broke up" sample from Step 2
- n_2 = number of individuals in your "stayed together" sample from Step 2
- $p_1 = p_2 = 0.30$
- *M* = number of chi-square tests you conducted in Step 2 (you do not need to count the chi-square test for trend).
- *G* = 1000

In 2-3 sentences, describe the results you obtained and state how these results inform your interpretation of the results you obtained in Step 2. How confident do you feel about the presence of statistically significant associations between remaining with partner and the variables in Step 2?

What you will submit:

1. A word document containing a brief written summary of the results you obtained in Step 2 above and the 2-3 sentences that you wrote following completion of Steps 3 and 4 (**1 single-spaced page maximum**). Within this word document, you will also include the tables produced in Step 2 and the figure produced in Step 3 (these do not have to fit within the 1 page).
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.

Projects are to be emailed to haralis@mednet.ucla.edu in an editable word document format no later than 5:00 PM on Friday, December 13th.