**Biostatistics 203A**

**Final Project**

Due: December 16, 2017

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 written report. The project is to be completed in 3 parts described below. You will be asked to submit three files:

1. A word document containing a brief written summary of your results (1-1.5 pages) and several tables and figures specifically requested in Parts 1 and 3.
2. A SAS syntax file containing the SAS macro described in Part 1.
3. An R Script containing the R function described in Part 3.

**Part 1**

Please download the csv file titled *Master\_Child\_SDQ.csv* from the course website. The file is meant to represent data collected on a sample of children enrolled in a multi-session behavioral and socioemotional intervention. Data consist of scores on 5 subscales of the Strengths and Difficulties Questionnaire, a self-report behavioral screening questionnaire for children and adolescents ages 4 through 17. Children were administered the questionnaire and scores on the subscales were calculated at each of 4 different time points (assessments) corresponding to the timeline of the intervention: baseline (prior to the intervention), exit (immediately following completion of the intervention), follow-up 1 (3 months after completion of the intervention), and follow-up 2 (9 months after completion of the intervention). Variables contained in the *Master\_Child\_SDQ.csv* file are as follows:

**SubjNum:** An identification number uniquely identifying each child

**SubjGender:** Child gender (“MALE” or “FEMALE”)

**SubjAge:** Child age in years (an integer between 4 and 17)

**Visit:** Assessment time point (“BASELINE”, “EXIT”, “FOLLOWUP1”, “FOLLOWUP2”)

**SDQ\_ESS:** Emotional Symptoms Subscale score (range: 0-10)

**SDQ\_CPS:** Conduct Problems Subscale score (range: 0-10)

**SDQ\_IHA:** Inattention-Hyperactivity Subscale score (range: 0-10)

**SDQ\_PPS:** Peer Problems Subscale score (range: 0-10)

**SDQ\_PSB:** Prosocial Behavior Subscale score (range: 0-10)

**RanNum:** Random number used to determine which data set you should use

All 5 of the subscales range from 0-10 (in theory). Higher scores indicate healthier behavior on the Prosocial Behavior Subscale. For the other 4 subscales, lower scores indicate less problematic behavior. The intervention is designed to target the behaviors measured in these subscales with the hope that a decrease in problematic behaviors and an increase in healthy behaviors will be observed from baseline to exit and will be sustained during the 9 months following completion of the intervention.

As with all longitudinal studies, there are two types of missing data: (1) missing data due to attrition, meaning children that decided to stop participating in the study at some point following the baseline assessment but prior to completion of the follow-up 2 assessment, and (2) random missed assessments. The data you will be working with contains a completed baseline assessment for all children, but you will observe NAs/missing values indicating missing assessment data for some children at some assessment time points.

The first thing you will need to do prior to completing any subsequent steps is to subset the *Master\_Child\_SDQ.csv* file so that you are working with only the records that have RanNum equal to your unique random number. This random number will differ for each student, but will be the same as the random number you used for the mid-quarter project. It can be found in the *Student\_Random\_Numbers.xls*x file on the course website.

Next, you will want to write SAS code to answer the following questions regarding this data set:

1. What number and percentage of children completed assessments at each of the 4 assessment time points (BASELINE, EXIT, FOLLOWUP1, and FOLLOWUP2)? The denominator for these percentages should be the total number of children. Please place these results in a table.
2. How many children completed only one assessment? How many children completed 2 assessments? 3? 4? Please provide the count and percentage for each number of assessments and place these results in a table.
3. What were the frequencies and percentages for each ‘combination’ of completed assessments present in the data? For instance, how many children completed only a BASELINE and a FOLLOWUP1 assessment? How many completed a BASELINE, EXIT and FOLLOWUP2 assessment? The denominator for these percentages should once again be the total number of children. Place these results in a table.
4. Calculate the mean and standard deviation for each SDQ subscale separately for each assessment time point. Present these numbers for the total sample of children and for subsamples corresponding to male and female children. Present these numbers in a table.
5. Identify the “last available” assessment time point for each child. By last available, we mean the chronologically last assessment time point with SDQ subscale responses not equal to NA/missing. Using scores from this last available time point, calculate a t-test to determine if statistically significant improvement was observed from baseline to last available assessment time point. Do this for each subscale and for the entire sample of children. Report the results of these t-tests in a table.

After having successfully completed the steps above, you should proceed to convert the SAS code you used to complete these steps into a SAS macro that can take 2 arguments corresponding to:

* The name of any SAS data set containing, at a minimum, the same variable names and values as found in the *Master\_Child\_SDQ.csv* data set
* The name of a two-level categorical variable that would exist in the provided data set and that would be used to in place of SubjGender in each of the results described above.

To test the effectiveness of your macro, you may choose to create a categorical variable from SubjAge (for instance, by categorizing it as ages 4-10 and 11-7) and call this macro using that variable as the second argument. You do not have to do this, it is just a suggestion. You will be submitting the macro and it will need to be executed during the grading of your assignment so please make sure it does not depend on any objects created locally on your machine.

**Part 2**

After having completed Part 1 in SAS, you will use R to complete the remainder of the project. In Part 2 your aim will be to simulate a data set with many of the same characteristics as the data set you used in Part 1. Do this by writing R code to accomplish the following:

* Create a data frame consisting of 4 rows for each of 100 subjects (the 4 rows corresponding to BASELINE, EXIT, FOLLOWUP1, and FOLLOWUP2)
* Sample Subject Gender with probability 0.50 of MALE and 0.50 of FEMALE
* Sample Subject Age from a uniform distribution with minimum 4 and maximum 17
* Sample baseline scores on each of the 5 SDQ subscales for each subject from a normal distribution with mean and standard deviation equal to the gender-specific sample means and sample standard deviations computed in Part 1. You must make sure that subjects assigned MALE gender in your simulated data set obtain a sampled value from the normal distribution with parameters specified based on the MALE sample in Part 1.
* Using the baseline scores, implement a “treatment effect” to obtain scores at each of the other 3 assessment time points. To do this, we will apply a multiplicative factor to the means being supplied as parameters to the normal distribution when sampling subscale scores at EXIT, FOLLOWUP1, and FOLLOWUP2 assessment time points. For instance, if you obtained the baseline score for a given subject in your simulated data set by sampling from a normal distribution with mean 1 and standard deviation 2, yielding a baseline score of 0.90 and the EXIT multiplicative factor is determined to be 0.65, you will obtain the EXIT score by sampling from a normal distribution with mean equal to 0.90\*0.65 = 0.585 and standard deviation 2 (do not change the standard deviation across time points). Use the following multiplicative factors:
  + EXIT = 0.65
  + FOLLOWUP1 = 0.80
  + FOLLOWUP2 = 1.05

These multiplicative factors are always applied to the mean parameter from the assessment immediately prior. Multiplicative factors will differ by assessment time point, but will not differ across subscales or genders. There is one exception, however, in that you will need to implement different multiplicative factors for the Prosocial Behavior Subscale, in which case the “treatment effect” should represent an increase in score from BASELINE, as opposed to a decrease. In this case you would want, for instance, to convert the 0.65 corresponding to a 35% decrease in mean score to a 1.35 (corresponding to a 35% increase).

* You should next apply each of the two missing data mechanisms:
  + For 10% of all subject assessment time points other than BASELINE, set values for all subscales equal to NA, such that random subjects will have random missing visits.
  + For each subject, select a last completed assessment time point and set all subscales equal to NA for subsequent assessment time points. Do this by sampling from the following probability distribution to determine which should be the last completed assessment for each subject:
    - BASELINE = 0.10
    - EXIT = 0.17
    - FOLLOWUP1 = 0.18
    - FOLLOWUP2 (all assessments completed) = 0.55

This means that 55% of subjects in your simulated data set will not have experienced attrition.

* Finally, for all subscales, if sampled scores are below zero, set them equal to zero. If sampled scores are greater than 10, set them equal to 10. All subscale scores should be rounded to the nearest integer, since integer scores are the only scores allowed on each of these subscales in practice.

To test whether your function works, you can always export the simulated data frame to a csv file, read it in to SAS, and execute the macro you developed in Part 1. The resulting statistics should be similar (although not exactly the same). This is just a suggestion and is not required.

**Part 3**

In completing Part 3, you will convert the R code you just wrote which generates a single simulated data set consisting of 400 rows reflecting 100 subjects, into an R function that can be repeatedly called. This R function should take the following arguments:

* Sample size (number of subjects, equal to 100 in Part 2)
* A proportion corresponding to the percentage in the first missing data mechanism (equal to 0.10 in Part 2)
* A vector corresponding to the proportions specified for the second missing data mechanism (c(0.10, 0.17, 0.18, 0.55) in Part 2).

You will also need to add some additional code to the code you wrote for Part 2. Specifically, each time the function is called, rather than one data set being generated the function should execute iteratively to generate 500 data sets. Each time a data set is generated, your code should conduct a t-test to evaluate the significance of the treatment effect for one of the subscales (you can select which subscale and it should be the same subscale each time) using the baseline assessment score and the last available assessment score as the two time points to compare using the t-test. Determine whether or not a significant treatment effect was identified for each of the 500 simulated data sets and have the R function you wrote return a single proportion corresponding to the percentage of times a significant treatment effect was identified. Significance should be evaluated at the 0.05 level.

Once your function is working such that on each call to the function you receive a single proportion, you can use this function to examine the impact of the following 3 things on the proportion of data sets in which a significant treatment effect was identified:

1. Changes in sample size
2. Changes in the percentage corresponding to the first missing data mechanism
3. Changes in the vector corresponding to the second missing data mechanism

For (1) and (2) above, use your function to vary the values and ultimately create two plots depicting the following:

* Sample size on the x-axis and percentage significant treatment effect on the y-axis
* Percentage corresponding to the first missing data mechanism on the x-axis and percentage significant treatment effect on the y-axis

You may use whatever type of plot you feel best describes these relationships. You may also select whatever range you feel is best for the x-axis values. Please be sure to label your plot appropriately. Include these 2 plots in the word document you submit. You will be submitting the function and it will need to be executed during the grading of your assignment so please make sure it does not depend on any objects created locally on your machine

**Projects are to be emailed to** [**haralis@mednet.ucla.edu**](mailto:haralis@mednet.ucla.edu) **in an editable word document format no later than 5:00 PM on Saturday, December 16th**.