August 5, 2022

Oregon Health & Science University
School of Medicine
3181 SW Sam Jackson Park Rd
Portland, OR 97239

Dear Colin Lipps,

Thank you for approaching me for some quick feedback on the statistical aspects of your study on sleep habits of medical students. I have reviewed the dataset you provided in forming the following commentary. I have made this preliminary analysis available to you in an R script appended to this correspondence. I recommend the use of R programming for this data analysis as it is free, easy to use, and there are bountiful online resources for exploratory and simple data analyses. You can download the software here (Download the RStudio IDE - RStudio), for which I recommend the integrated development environment RStudio. Some lecture slides on R programming can be found here courtesy of my colleague: STAT 302 (bryandmartin.github.io).

Preliminary Analysis

- Your study includes 20 subjects who participate in as many as 27 rotational assignments. You have collected data concerning their daily sleep behaviors and recorded these over a 1-2 year time period. You are most interested in how rotational assignments impact the sleep behaviors of the medical students.
- Your data is well-balanced with respect to the days of the week. Each weekday has nearly 700 observations.
- Your data is not well-balanced with respect to the rotational assignments. Some rotations have many observations (more than 100) and some rotations have few observations (less than 100). This dataset attribute will impact the standard errors / interval estimates for rotation-specific population mean sleep behaviors; namely, there will be more uncertainty with less data.
- You do have some covariates (columns in your dataset) that have a high percentage of missing values. These are mostly not the columns you highlighted in yellow as of principal interest. For the time being, I recommend studying columns “RHR” to “Sleep Debt” and ignoring the 13-22 missing values. If the other columns are of interest, please reach out to discuss missing data and imputation strategies.
- As exploratory data analysis, I plotted the sleep variables over time/date. These time series plots appear to have no general trend. In exploratory model building, I also find little evidence for monthly or seasonal effects. This will simplify your data analysis since you should not need to

adjust for temporal effects. You can find an example time series plot for "Hours of Sleep" at the end of this letter.

- As exploratory data analysis, I plotted box-and-whisker plots for the sleep variables per weekday. From these plots, I see little evidence of a weekday trend in sleep behaviors. In exploratory model building, I sometimes see 1 or 2 weekdays have a noticeable effect for some sleep variables. Nevertheless, I recommend you keep the data analysis simple for now and ignore weekday.
- As exploratory data analysis, I plotted box-and-whisker plots for the sleep variables by rotation code. I see evidence for some heterogeneity in the sleep variables based on the rotational assignments. You can find an example box-and-whisker plot for "Hours of Sleep" at the end of this letter.
- Your dataset has a longitudinal component to it in that the observations are correlated by medical student. Each medical student has distinct sleep characteristics. I recommend a (generalized) linear mixed model approach to data analysis with a random intercept term for each medical student. You can find details about random intercept models here (Random intercept models | Centre for Multilevel Modelling | University of Bristol) , here (Multilevel model - Wikipedia), and here (Mixed Models: Testing Significance of Effects (wisc.edu)). While I acknowledge that your research goals focus on the population mean effects of rotational assignments on medical student sleep variables, I view it as important to include the subject-specific heterogeneity in statistical modeling. Initially, it does seem that a variance component hypothesis test for the random intercept would conclude that a random intercept is statistically significant.
- Your dataset has some sleep variables that appear to be percentages, bounded by the interval [0,100]. Linear mixed models assume that the dependent/response variable is normally distributed. You may need to consider some simple modifications to the analyze of the percentage sleep variables. "HRV" and "Hours of Sleep" appear straightforward to analyze without such a modification. Please reach out to discuss this point for the other sleep variables.
- You mentioned an ANOVA test to me. Without better understanding if there is a specific research question that informs a hypothesis test, I cannot recommend statistical hypothesis testing. I opine that there are statistics and data visualization techniques available to present the heterogeneity and differences among sleep variables by rotational assignment, without defaulting to the language of *p*-values, significance, etc. For example, we can see in box-and-whisker plots and interval estimates for the fixed effects (population mean effects of rotational assignments) this heterogeneity. Moreover, inference and hypothesis testing for linear mixed models (LMMs) is an advanced topic still of research interest to many statisticians, which would mean you may require considerable statistical help to interpret these.
- Standard practice is to form linear models with a fixed effect intercept term. In this setting, the interpretation of model effects (beta terms) on categorical variables is with respect to a base/reference category (by default rotation code 1). You have categorical variables as independent variables, the rotation codes. I recommend instead a linear model without the fixed

effect intercept term; in this setting, the model effects (beta terms) provide the population mean sleep behavior for each rotation code. I implemented this model and computed naïve 95% confidence interval estimates for “Hours of Sleep”. These are appended to the end of this letter.

Follow-up Questions

- Do you have a specific research question? If so, could you share this with me and we could discuss if it is amenable to statistical hypothesis testing and/or discuss this further?
- Who is the intended audience of this data analysis? How does this data analysis fit into the broader context of your research?
- Could some of the rotation codes be combined? For instance, it appears that some rotations are similar, but at different sites.

I hope that this correspondence has been helpful and thought-provoking as you approach your continuing data analysis. Please reach out if you have any clarifying questions to discuss. As well, if requested, I can make some of the appended figures into a production-quality form or demonstrate code to do as much.

Sincerely,

Seth D Temple

PhD Student

NDSEG Fellow

NIH Trainee in Statistical Genetics

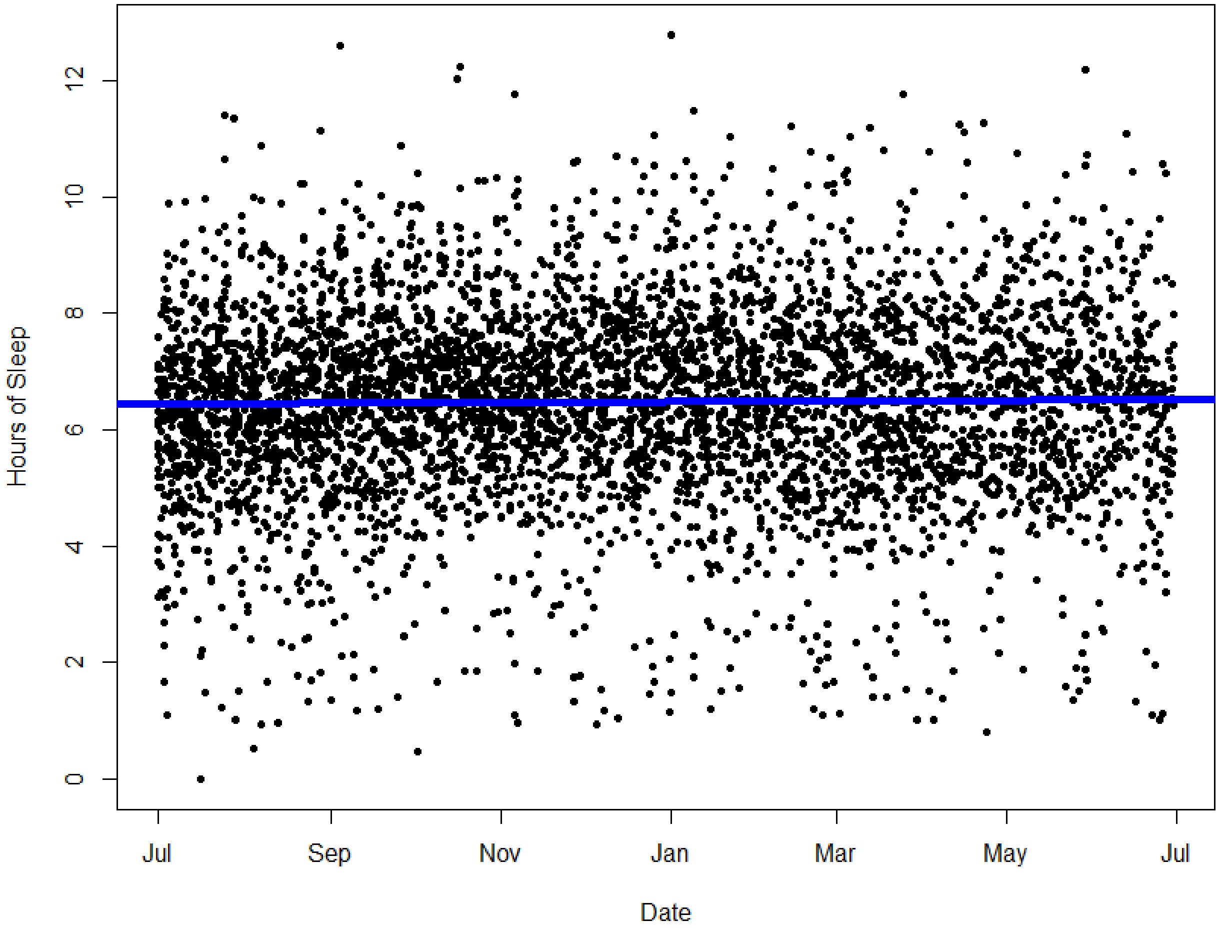

Figure 1 : Time series plot for hours of sleep

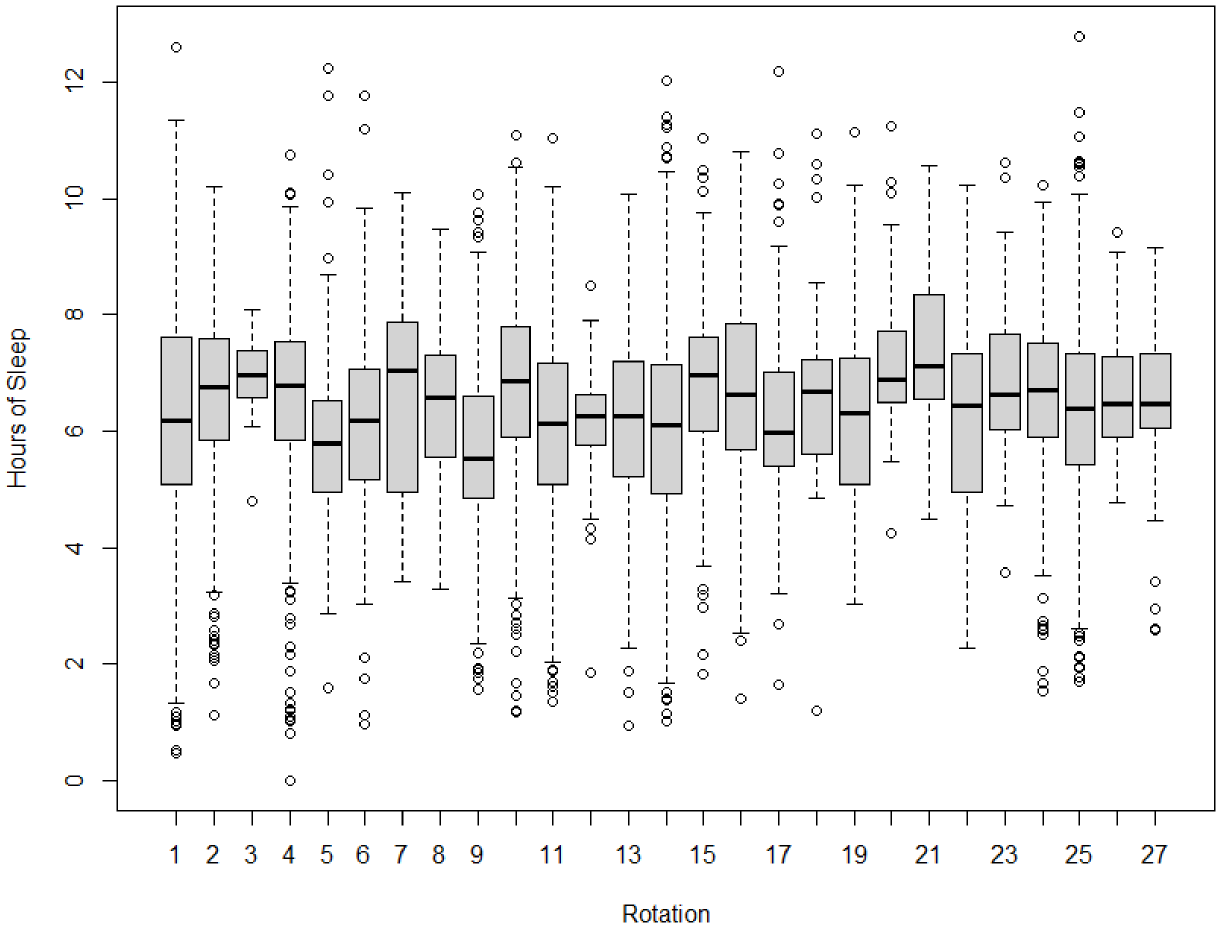

Figure 2 : Box-and-whisker plots for hours of sleep by rotation code

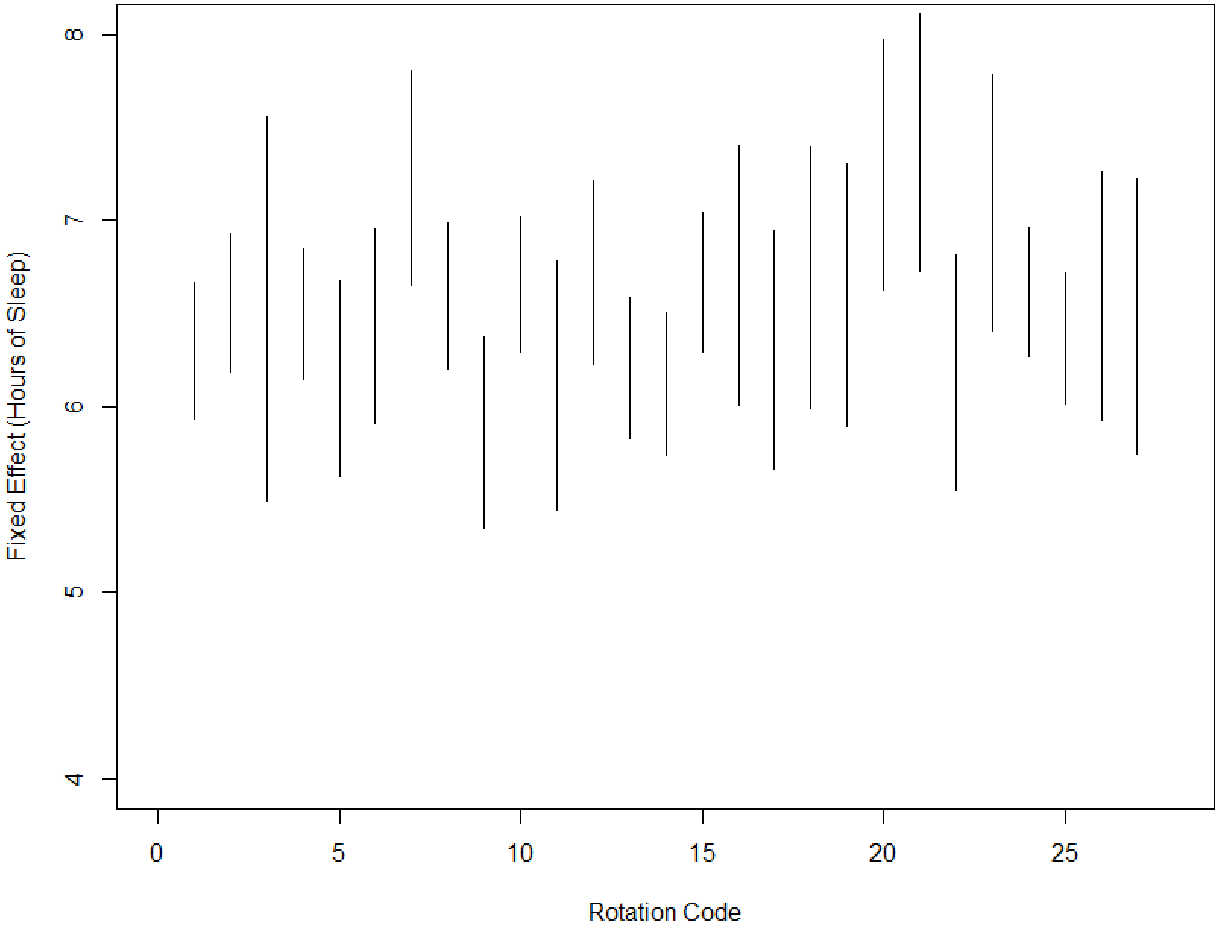

Figure 3 : Interval estimates for rotation code population mean effect on hours of sleep