Biostatistics 140.656

Lab 0

**Topics:**

* Data management for multi-level data
* Computation of descriptive statistics for level 2 variables
* Computation of descriptive statistics for level 1 variables

**Learning Objectives:**

Students who successfully complete this lab will be able to:

* Compute descriptive statistics for the number of level-1 units within level-2 units
* Compute descriptive statistics for level-2 variables
* Compute descriptive statistics for level-1 variables across all level-1 units
* Define a calculated variable and compute descriptive statistics for calculated level-2 variables

**Scientific Background**:

In Homework 1, you will be analyzing a cross-sectional study of high school mathematics achievement from the High School and Beyond (HS&B) study conducted within the National Education Longitudinal Studies (NELS) program of the National Center for Education Statistics (NCES). The NELS was established to study the educational, vocational, and personal development of young people beginning with their elementary or high school years, and following them over time as they begin to take on adult roles and responsibilities. Thus far, the NELS program consists of five major studies: the National Longitudinal Study of the High School Class of 1972 (NLS-72), High School and Beyond (HS&B), the National Education Longitudinal Study of 1988(NELS:88), the Education Longitudinal Study of 2002 (ELS:2002), and the High School Longitudinal Study of 2009 (HSLS:09).

The HS&B survey included two cohorts: the 1980 senior class, and the 1980 sophomore class. Both cohorts were surveyed every two years through 1986, and the 1980 sophomore class was also surveyed again in 1992.

We have available data from one of the assessments for 7042 students within 156 schools.

The study variables include:

Level 1: student

mathach: a measure of mathematics achievement

minority: dummy variable for student being non-white

female: dummy variable for student being female

ses: socioeconomic status (SES) based on parental education, occupation and income (z-score)

Level 2: school

schoolid: school identified

sector: dummy variable for a school being Catholic

pracad: proportion of students in the academic track

disclaim: scale measuring disciplinary climate

himinty: dummy variable for more than 40% minority enrollment

size: number of students enrolled at the school

newid: rescaled school identifier, counts 1 to 156 (we created this for you)

**Lab Exercise:**

1. Data structure:

Multi-level data are most commonly arranged in the “long” format; e.g. each row of the dataset represents a student within a particular school. Create two new variables in your dataset; a) the number of students within each school and b) a within school student ID that counts from 1 to the number of students within each school.

View a few variables from the first school (newid = 1) which has 47 students; based on the output, confirm that the variables “N” and “sector” are variables defined at the school level and that “n” and “ses” are defined at the student level.

**Stata:**

use “hsb\_data”, clear

bys newid: gen N = \_N

bys newid: gen n = \_n

list newid N n ses sector in 1/47

**R:**

data <- read.table(“hsb\_data.csv”,header=T,sep=”,”)

N <- unlist(tapply(data$schoolid,data$schoolid,length))

data$N <- rep(N,N)

data$n <- unlist(tapply(data$schoolid,data$schoolid,FUN=function(x) seq(1,length(x))))

data[1:47,c(“newid”,”N”,”n”,”ses”,”sector”)]

1. School-level variables:

It is important to understand the characteristics of the schools, i.e. the level-2 units. Create a table summarizing the school-level variables in the HS&B dataset.

In addition, it is useful to understand how the school-level characteristics are distributed across the entire population of students. E.g. if 45% of the schools are Catholic schools, do 45% of the students in the dataset attend Catholic schools? Append this additional student level description of the school-level variables to your summary table.

1. Student-level variables:

It is important to describe the characteristics of the students that define the population. Create a table summarizing the available student-level, i.e. level-1, characteristics.

In addition, it is important to maintain the hierarchical structure of the data in describing student-level characteristics. Another way to describe this is to consider that schools can vary in their student composition. Calculate additional statistics that allow you to describe how the student composition varies across the schools.