Biostatistics 140.656
Lab 0 Solution

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")]
```


2. 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.

   SOLUTION: Table 1 below displays one example of a table that summarizes school characteristics at both the school-level and student-level. NOTE: 45% of the schools are Catholic schools where 50% of the students in the sample come from catholic schools.

   See the Stata and R code for how to complete the computations.

Table 1: School characteristics summarized at the school level and for the students included in the sample.

Variable	Schools (n = 156)	Students (n = 7042)
Catholic school, % (n)	45 (70)	50 (3543)
More than 40% minority enrollment, % (n)	26 (40)	27 (1869)
Proportion of students on the academic track, mean (SD)	.52 (0.25)	0.54 (0.25)
School disciplinary climate, mean (SD)	-0.05 (0.95)	-0.15 (0.93)
Number of students enrolled in the school, mean (SD)	1073 (613)	1037 (590)

3. 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.

   SOLUTION: Table 2 below displays one example of a table that summarized the characteristics of the students in our sample and then also compares the student characteristics across the schools (i.e. composition of the students).

   The student average math achievement score is 12.9. However the average math achievement score within schools varies with 50% of the schools having average math achievement score between 10.9 and 14.7.

   Fifty three percent of the students are female. However, 18 and 19 schools have all male and female students in the sample, respectively. The percentage of female students in the school samples vary with 50% of the schools having 44 to 61% female students.

See the Stata and R code for computations required to fill in Table 2.

Table 2: Descriptive statistics for the student characteristics are displayed. In addition, descriptive statistics are displayed to demonstrate how the student composition varies across the schools.

Variable	Students (n = 7042)	Schools[1] (n = 156)
Math Achievement, mean (SD)	12.9 (6.8)	13.0 (10.9, 14.7)
Non-white, % (n)	26 (1831)	14 (4, 40)
Female, % (n)	53 (3714)	53 (44, 61)[2]
SES, mean (SD)	0.0 (0.8)	0.0 (-0.3, 0.3)

[1] Values represent the median (Q1, Q3) of the mean math achievement and SES or the proportion of non-white and female students across schools.
[2] 18 and 19 schools are represented by only male and female students, respectively.

In addition, the figures below display the distribution of the student characteristics across the schools.