Basics of Stata

Aim of the lab

- Open and describe the dataset
- Summary statistics
- Table of counts and summary statistics
- Comparisons of proportions and means
- Generate, replace, and recoding variables
- Plot the risk of the disease by exposure categories using a bar chart

Open and describe the dataset

1. Open the dataset **hyponatremia.dta** located on the webpage of the course (help use).

   Get familiar with the information contained in the dataset (help describe, help codebook). How many subjects and variables are in the dataset? And what kind of variables (categorical, continuous)?

2. What is the range of **id** and how many unique values are there? Can you see duplicated values for id?

3. How many missing values have the variables weight change, body-mass index, and race duration? What is the proportion of missing data for each of the three variables?

4. Identify runners with critical hyponatremia (sodium concentration below 120 mmol per liter). For example, list the variable **id** and **na** (help list, help browse) for those runners. How many are they? What is the proportion of runners with critical hyponatremia?

Presentation of summary statistics

5. The variable **na** contains the serum sodium concentration (continuous response) expressed in mmol per liter. Describe the distribution of the response variable **na** (help summarize, help histogram, help graph box).
   What is the mean, standard deviation, and range of sodium concentration in the sample of runners?

6. What is the mean and standard deviation of race duration among male and female runners?

7. What is the mean and standard deviation of race duration among cases and non-cases of hyponatremia?

8. Produce a table containing min, median, and max of sodium concentration by categories of weight gain.

Comparisons of proportions and means

9. What is the proportion of runners with self-reported frequency of voiding during race 3 times or more? How this proportion varies of among cases and non-cases of hyponatremia? Which test can you use to compare proportions? What do you conclude about the association between voiding 3 times or more during race and the risk of hyponatremia? Motivate your answer.

10. Is there any statistical evidence of difference in the mean race duration comparing the populations of cases and non-cases of hyponatremia? What is the difference? Motivate your answer.

Generating and replacing values of a variable

11. To investigate the association between running time and the risk of hyponatremia you are asked to categorize race duration in three levels (<3:30, 3:30-4:00, >4:00 hours). Check the distribution of the categorized variable. What is the percentage of runners above 4 hours?

12. Next, similarly to Table 2 of the NEJM paper present the percentage distribution of race duration by case status (help tabulate). What is the p-value of the association between race duration in categories and the risk of hyponatremia?

Graphs

13. Suppose you want to publish a figure describing the relationship between race duration and the risk of hyponatremia (help graph bar). Something similar to Figure 1 of the paper. Create a good-looking two-way plot suitable for publication controlling various element of the graph.