BIOSTAT 702: Exercise 1.1

Describing Participant Selection into A Research Study

Fall 2025

Contents

How to Do This Exercise

We recommend that you read this entire document prior to answering any of the questions. If anything is unclear please ask for help from the instructors or TAs before getting started. You are also allowed to ask for help from the instructors or TAs while you are working on the assignment. You may collaborate with your classmates on this assignment—in fact, we encourage this–and use any technology resources available to you, including Internet searches, generative AI tools, etc. However, if you collaborate with others on this assignment please be aware that *you must submit answers to the questions written in your own words. This means that you should not quote phrases from other sources, including AI tools, even with proper attribution.* Although quoting with proper attribution is good scholarly practice, it will be considered failure to follow the instructions for this assignment and you will be asked to revise and resubmit your answer. In this eventuality, points may be deducted in accordance with the grading rubric for this assignment as described below. Finally, you do not need to cite sources that you used to answer the questions for this assignment.

Grading Rubric

The assignment is worth 20 points (4 points per sub-question, as specified). The points for each question are awarded as follows: 3 points for answering all parts of the question and following directions, and 1 point for a correct answer. Partial credit may be awarded at the instructor's discretion.

Resources

The following resources on Canvas will be helpful for answering the questions for this exercise.

1. The Short Report that explains the STROBE statement
2. The 'explanation and elaboration' paper on the STROBE statement, which can serve as a reference if you need it while answering the questions for this exercise
3. The ultrarunning manuscript by Samtleben
4. The ultrarunning data dictionary for the study by Samtleben
5. The ultrarunning dataset for the study by Samtleben

Question 1 (4 points)

For the ulratrunning paper by Samtleben, find the first 10 elements of the STROBE statement and fill in the checklist (you can ignore items 11-22 of the checklist). Some things to be aware of as you do this:

- You should be aware that some of the items on the checklist might not appear in the paper and you should note this on the checklist.
- You will also likely find that some of the statistical analyses described in the paper might be unfamiliar to you. Don't let this be a concern for now; it should not prevent you from completing the STROBE checklist.
- In later exercises will we conduct a simpler analysis than what Samtleben discusses in the paper. Our analysis will be a simple linear regression (SLR) with emotional intelligence as the predictor and best ultra-running time as the outcome (see Question 3).

Question 2 (4 points)

As mentioned in class, we have 3 overlapping groups to consider: (1) the target population to whom we would like to generalize the results of the study; (2) those who enrolled in the study (n=288); and (3) those with non-missing values of the predictor and the outcome who were analyzed (N=125). Samtleben doesn't precisely define the target population, but does discuss it within the context of differences between the target and sample populations.

1. As precisely as possible, what do you believe the target population to be based on Samtleben's description?
2. What might you ask the investigator to clarify your definition of the target population?

Question 3 (12 points)

We will eventually conduct a simple linear regression with emotional intelligence as the predictor (independent variable; teique_sf) and best ultra-running time as the outcome (dependent variable; pb100k_dec). Some participants might be dropped from our analysis because of missing values on either of these variables. The term "selection bias" describes systematic differences between participants who were enrolled vs. those who were analyzed. Your task is to assess the degree of bias, if any, caused by dropping observations with missing values. Do this informally (that is, without performing statistical tests and generating p-values) using the Visualize, Analyze, Interpret (VAI) framework.

1. Visualize (4 points): Perform a descriptive analysis of the two variables we are interested in studying for this analysis. Assuming we will remove observations missing either of these variables, how many observations will be removed? What will be the size of the sample and the size of the analytic dataset? Note: This number might differ from the paper, due to different variables used.

2. Analyze (4 points): Check for systemic differences between those in the analytic dataset and those excluded from the analytic dataset. Look for potential systemic differences in the following variables: pb_surface, pb_elev, avg_km, steu_b, stem_b. To do this, create a Table 1 stratified by inclusion in the analysis, with standardized mean differences.

3. Interpret (4 points): Based on the Table 1 you created, do you notice any systemic differences in the other variables that lead you to believe there may be bias in your analysis? Are there other sources of bias that you may not be able to interpret just by looking at this Table 1?