BIOSTAT 702: Exercise 1.1

Describing Participant Selection into A Research Study

August 13, 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).

See the attached checklist with items 1-10 completed.

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?

   The intended target population is likely to be English-speaking adult men and women who regularly participate in 100km ultra races.

2. What might you ask the investigator to clarify your definition of the target population?

   Here are some things to consider (not an exhaustive list): What experience level? Where do they live/race? What about other kinds of ultra races (shorter/longer)?

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.

```
# The target population is not well defined or enumerable. For now we consider
# the target population to be English-speaking adult men and women who regularly
# participate in 100km ultra races.
#
# Although we can't describe the target population in detail we can describe the
# enrolled and analyzed participants and compare them to assess any systematic
# differences.
#
# The first step is to load the data and inspect it to make sure the structure
# of the data set meets our expectations, and that the data are not peculiar.
#
# Load the data
ultra <- read.csv(here::here("Datasets/Ultrarunning/ultrarunning.csv"))

# See what the structure of the data set looks like
str(ultra)
```

```
## 'data.frame':    288 obs. of  10 variables:
##  $ age        : num  49 36 29 26 40 27 29 36 36 37 ...
##  $ sex        : num  2 2 2 2 1 2 1 2 1 2 ...
##  $ pb_surface : int  1 3 1 1 1 1 3 1 3 1 ...
##  $ pb_elev    : num  2018 631 1524 3657 4420 ...
##  $ pb100k_time: chr  "14:00:00" "07:36:12" "14:12:00" "14:20:00" ...
##  $ pb100k_dec : num  14 7.6 14.2 14.3 17 ...
##  $ avg_km     : num  70 110 80 110 70 105 125 55 50 30 ...
##  $ teique_sf  : num  NA 5.73 5.33 5.33 5.33 5.23 5.97 5.37 5.2 5.67 ...
##  $ steu_b     : num  15 14 15 14 13 15 15 13 14 13 ...
##  $ stem_b     : num  NA 13.5 10.8 10.6 13.6 ...
```

```
head(ultra)
```

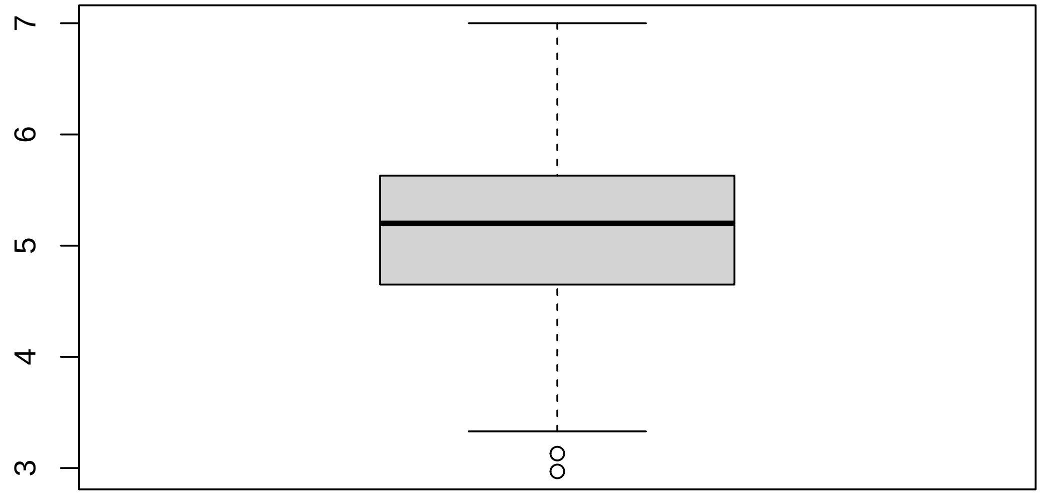
```
##   age sex pb_surface pb_elev pb100k_time pb100k_dec avg_km teique_sf steu_b
## 1  49   2          1    2018    14:00:00      14.00     70        NA     15
## 2  36   2          3     631    07:36:12       7.60    110      5.73     14
## 3  29   2          1    1524    14:12:00      14.20     80      5.33     15
## 4  26   2          1    3657    14:20:00      14.33    110      5.33     14
## 5  40   1          1    4420    17:00:00      17.00     70      5.33     13
## 6  27   2          1    1372    12:00:00      12.00    105      5.23     15
##   stem_b
## 1     NA
## 2  13.50
## 3  10.75
## 4  10.58
## 5  13.58
## 6  12.33
```

```
# Descriptive analysis of the independent variable
#
# Per the data dictionary: "variable 8 is the Emotional Intelligence score,
# defined as the average of the 30 items of the Trait Emotional Intelligence
# Questionnaire Short Form"

summary(ultra$teique_sf)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.970   4.650   5.200   5.125   5.630   7.000      73
```

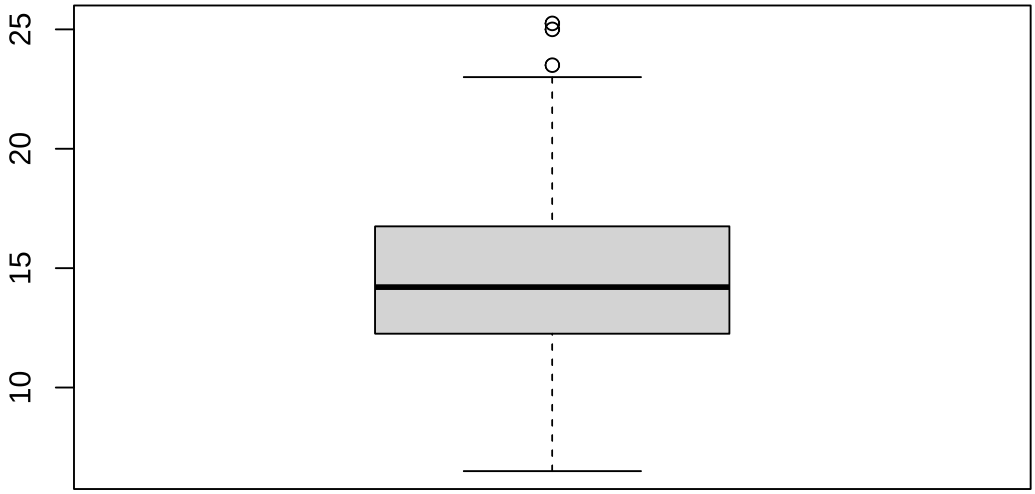
```
# We see immediately there are 73 missing values. Lets look at the distribution.
boxplot(ultra$teique_sf)
```



```
# Descriptive analysis of the dependent variable.
# There are two versions. One is given in minutes:hours:seconds, which is a
# character variable. The second variable, which is the one we want to use,
# is the number of hours.
summary(ultra$pb100k_dec)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.50   12.26   14.21   14.73   16.75   25.25
```

```
boxplot(ultra$pb100k_dec)
```



The descriptive analysis above shows us that there are no missing values for the dependent variable, pb100k_dec,but there are missing values for the independent variable, teique_sf. When we perform a simple linear regression we will lose the 73 participants for whom the independent variable is missing. Therefore, our task is to describe differences in baseline characteristics between the 288 who are enrolled and the 215 who would be included in our simple linear regression.As the descriptive analysis illustrates, our task of comparing these two groups of patients is somewhat complicated by the fact that there are missing data for some of the variables we'd like to compare between the groups.

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.

```
# First, we need to create an indicator variable that represents who will
# be excluded from the simple linear regression vs. who will be included. Note
# that we write the program code to exclude based on missingness in either the
# dependent or independent variable even though we know in this case that the
# missingness is only in the independent variable. It is good practice to
# program for all contingencies in case, for example, the dataset is updated
# at a later date.
ultra$excluded <- ifelse( is.na(ultra$teique_sf) | is.na(ultra$pb100k_dec), 1, 0)
ultra$excludedF <- factor(ultra$excluded,levels=c(0,1),labels=c("Included","Excluded"))
table(ultra$excludedF)
```

```
##
## Included Excluded
##      215       73
```

```r
# The variables we are concerned with for our assessment of bias are age, sex,
# pb_surface, pb_elev, avg_km, steu_b, and stem_b.
table(ultra$sex,useNA='always')
```

```
##
##    1    2 <NA>
##   92  195    1
```

```r
table(ultra$pb_surface,useNA='always')
```

```
##
##    1    2    3    4 <NA>
##  210    4   38   36    0
```

```r
summary(ultra$pb_elev)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     792    2400    2681    3962   13000      24
```

```r
summary(ultra$avg_km)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.00   60.00   72.00   74.44   90.00  160.00       7
```

```r
summary(ultra$steu_b)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    6.00   11.00   13.00   12.66   14.00   18.00      76
```

```r
summary(ultra$stem_b)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    4.67   10.33   11.54   11.42   12.81   15.08     126
```

```r
# A couple observations are key here: 1) The variables we want to compare
# between enrolled and analyzable participants themselves have missingness; and
# 2) Two of these variables are categorical (sex and pb_surface) and we should
# define factor variables so that labels are attached to the categories.
# We will ignore the first problem for now, and proceed to solve the second
# issue by creating factor variables.
ultra$sexF <- factor (ultra$sex,levels=c(1,2),labels=c("Male","Female"))
ultra$pb_surfaceF <- factor(ultra$pb_surface,levels=c(1,2,3,4),
                            labels=c("Trail","Track","Road","Mix of all three"))
table(ultra$sex,ultra$sexF,useNA='always')
```

```
##
##        Male Female <NA>
##   1      92      0    0
##   2       0    195    0
##   <NA>    0      0    1
```

```
table(ultra$pb_surface,ultra$pb_surfaceF,useNA='always')
```

```
##
##        Trail Track Road Mix of all three <NA>
##   1      210     0    0                0    0
##   2        0     4    0                0    0
##   3        0     0   38                0    0
##   4        0     0    0               36    0
##   <NA>     0     0    0                0    0
```

```
# Now we will use the tableone package to create a table that compares
# these two groups of participants.
library(tableone)
```

```
## Warning: package 'tableone' was built under R version 4.4.1
```

```
vars <- c("age","sexF","pb_surfaceF","pb_elev","avg_km","steu_b","stem_b")
comparisonTable <- CreateTableOne(vars = vars, strata = "excludedF",
                                  data = ultra, test=FALSE)
print(comparisonTable, smd=TRUE, missing=TRUE)
```

```
##                        Stratified by excludedF
##                         Included          Excluded          SMD    Missing
##   n                         215                73
##   age (mean (SD))         39.78 (10.50)     39.60 (10.70)    0.017  0.7
##   sexF = Female (%)         147 (68.7)         48 (65.8)     0.063  0.3
##   pb_surfaceF (%)                                            0.255  0.0
##      Trail                  156 (72.6)         54 (74.0)
##      Track                    4 ( 1.9)          0 ( 0.0)
##      Road                    26 (12.1)         12 (16.4)
##      Mix of all three        29 (13.5)          7 ( 9.6)
##   pb_elev (mean (SD)) 2568.39 (2047.55) 3010.53 (2532.87)  0.192  8.3
##   avg_km (mean (SD))      74.07 (24.26)     75.55 (27.48)    0.057  2.4
##   steu_b (mean (SD))      12.58 (2.50)      13.18 (2.18)     0.257 26.4
##   stem_b (mean (SD))      11.42 (1.79)      11.36 (2.22)     0.034 43.8
```

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?

   In this step we have produced a table that compares the distribution of baseline variables between the set of participants who will be included in our simple linear regression and the set of participants who will be excluded. Some of these variables are missing for a large percentage of the participants and so are not helpful. For example, 26.4% and 43.8% of participants are missing data on emotional understanding (steu_b) and situational management (stem_b). Missingess is less than 10% in other variables, however, which makes us more comfortable using them as a basis for comparing these two groups of participants. The only factor for which the standardized mean difference is higher than 0.2 is the type of surface that participants had their personal best running time on (pb_surface). By looking at the descriptive statistics we can tell that the participants we we will analyze are more likely to have achieved their personal best running time on a mixture of trail, track, and road surfaces. This might suggest the runners in our anlaysis have a higher skill level than runners we excluded. This could limit the generalizability of our results to everyone in the study (an issue of internal validity) as well as to the target population (an issue of external validity). From this Table 1, we cannot tell how well our sample or analytic dataset represents the target population (sampling bias) or if there is any kind of information bias.