Exercise 3: Initial Data Checks

Author: Leilani Madrigal

Department of Kinesiology

California State University, Long Beach

Long Beach, CA 90840

Email: [Leilani.madrigal@csulb.edu](mailto:Leilani.madrigal@csulb.edu)

**Note to the Instructor**: The data set used in this exercise (**Sport Injury and Anxiety**) is data that was collected and put together for the purpose of this exercise module. This exercise uses FREQUENCIES and SELECT CASES in SPSS to identify incorrect data entry.  A good reference on using SPSS is SPSS for Windows Version 23.0 A Basic Tutorial by Linda Fiddler, John Korey, Edward Nelson (Editor), and Elizabeth Nelson.  You have permission to use this exercise and to revise it to fit your needs.  Please send a copy of any revision to the author. Included with this exercise (as separate files) are the SPSS file carry out the exercise (SPSS file\_data), the Online Questionnaire (Online Questionnaire PDF), answers to the exercise (exercise 3\_answers), and SPSS output for the exercises (exercise 3\_output.sav). Please contact the author for additional information.

Attached are files for this exercise:

* SPSS file\_data (SPSS file)
* Online Questionnaire (PDF)
* Exercise 3\_answers (MS word document)

Exercise 3\_Output (SPSS output file)

**Goals of Exercise**

The goal of this exercise is to learn to check that data was entered correctly by examining frequencies and visual inspections of all variables in the newly constructed SPSS file.

**Checking Frequencies**

Now that all data has been entered from the excel file to your newly constructed SPSS file (use **SPSS file\_data)**, it is time to make sure all values were correctly entered. This is a necessary step to ensure proper data management. First, let’s review what the ranges of frequencies should be for the different variables, based on the online questionnaire.

***Review of Demographic Information***: Based on the online questionnaire, demographic variables included age, year in school, geographic location, gender, ethnicity, sport, season, NCAA division, injury history, currently injured. For variables that had associated values (Ethnicity was 1-7), Sport was (1-8), etc. We now must check that all our demographic variables are within the expected values/ranges that were assessed.

To do this in SPSS, once the file is open, click on “Analyze” 🡪 “Descriptive Statistics” 🡪 “Frequencies”. Then click on each variable you’d like to examine by click on it and then the arrow button to move the variable to the “Variable(s)” box. Once all variables that you want to examine are in that box, click on “Statistics” which will prompt a new dialogue box to open. Within this box, make sure you click on *mean, std. deviatio*n, as well as *minimum* and *maximum*. Then click “continue” and then “ok”.

You will see an output displayed of the frequencies for those variables. Below is an example output for the variable “sport”.

|  |  |  |
| --- | --- | --- |
| **Statistics** | | |
| Sport | | |
| N | Valid | 131 |
| Missing | 2 |
| Mean | | 10.19 |
| Std. Deviation | | 4.852 |
| Minimum | | 1 |
| Maximum | | 22 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sport** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Basketball | 5 | 3.8 | 3.8 | 3.8 |
| Beach Volleyball | 15 | 11.3 | 11.5 | 15.3 |
| Field Hockey | 8 | 6.0 | 6.1 | 21.4 |
| Golf | 1 | .8 | .8 | 22.1 |
| Lacrosse | 44 | 33.1 | 33.6 | 55.7 |
| Soccer | 26 | 19.5 | 19.8 | 75.6 |
| Swimming | 12 | 9.0 | 9.2 | 84.7 |
| Track and Field | 3 | 2.3 | 2.3 | 87.0 |
| Volleyball | 5 | 3.8 | 3.8 | 90.8 |
| 19 | 7 | 5.3 | 5.3 | 96.2 |
| 20 | 1 | .8 | .8 | 96.9 |
| 21 | 3 | 2.3 | 2.3 | 99.2 |
| 22 | 1 | .8 | .8 | 100.0 |
| Total | 131 | 98.5 | 100.0 |  |
| Missing | 999 | 1 | .8 |  |  |
| System | 1 | .8 |  |  |
| Total | 2 | 1.5 |  |  |
| Total | | 133 | 100.0 |  |  |

Notice on the Statistics table are the mean, std. deviation, minimum, and maximum for each variable. As you scroll down you will see the frequencies for that variable. Notice that there is a value of 19, 20, 21 and 22 but no associated label. This could mean that a number was entered incorrectly, or there may have been additional sports that were not correctly labelled. You would need to go back to the “online questionnaire” and see if there were in fact 12 sports or if these were errors in data entry.

**Exercise 1:** Check the frequencies for all demographic variables: age, year in school, geographic location, gender, ethnicity, sport, season, NCAA division, injury history, currently injured. For this step, if you do find an error just report what variable had a value out of range.

***Review of Scales:*** Looking at the remaining questions of the online questionnaire, you will see that the rest of the items were a collection of different measures. For each measure a corresponding Likert-type range exists. For example, the TSK1-TSK17 has participants rate their degree of agreeableness on a scale of 1-5, with 1= strongly disagree and 5= strongly agree. So when examining frequencies for variables TSK1- TSK17, all values should range between 1-5.

**Exercise 2**: Check the frequencies for the remaining items of the online survey. For this step, just report what variable had a value out of range.

Did you catch any data inputting errors? It is common to catch at least 1-2 at this step, this is why it is so important.

**Selecting Cases to Identify an Incorrect Data Entry**

For smaller data sets, it may be easy to identify where the problematic item is within the data set. This can be a visual scan to examine the value that was out of range for the variable you identified by running frequencies. However, there are simpler, less eye straining ways of finding the rogue data entry. In SPSS, we can use the “select cases” function to help identify these variables.

Once you’ve identified a problematic variable, let’s take “sport”, we need to see which participant had an incorrect value inserted. So in SPSS, click on “Data”🡪 “Select Cases”. A dialogue box will open, click on “If condition is satisfied” 🡪 “If…” A new dialogue box opens and this is where you can specific what conditions must be bet for the data to be examined. For example, if I wanted to just examine responses from female participants, I could select “gender” and click on the arrow to move it to the open box, and indicate Gender = 1 and click “continue” 🡪 “ok”. This will prompt SPSS to only use data sets where gender was equal to 1 (which I coded as 1 = female). You can see on the “data view” that cases are all crossed out if gender = 2 (males).

Using select cases to identify an error in data entry is another great function. So going back to our example, if I move “sport” into the open box of my select cases dialogue box. I need to specify what the conditions will be. Since I know a value greater than 18 was out of range of the “sport” variable. I will indicate Sport > 18 . Then click “continue” 🡪 “ok” This way, all cases in which that is not true will be identified. However, this may be still hard to determine which ID (or case) is the problematic (especially if the data set has a large number of participants/responses). So another trick, is to click “File” 🡪 “New” 🡪 “Syntax”. This prompts a new window to pop up. For the ease of understanding I will not get into all the different things you can do with syntax files, but for the purposes of using this to help clean data, the only command for you to use will be “List ID.” It’s very important that you put a period after ID. This tells SPSS you are done giving a command. After you write that in, you will notice “List” turned blue. Now click on the Green Arrow, or you can click on “Run” 🡪 ALL. Go to your output window and you will see the associated ID’s that meet that criteria. For this example, this was ID’s 19,20,23,24,47,56,102,107,109,118,128,129. The next step would be to go back to your data sets for these ID #’s and input the correct value (or 999) if that is the more appropriate option.

**Exercise 3:** For all the variables you identified in exercises 1 and 2, find the cases/IDs in which the error occurred. Just list the IDs and variables.

NOTE: to truly correct each of the IDs you would need to go back to the raw data to see what the value should be (if you did a survey in person, this may mean going back to the hard copy/paper version of the survey to see what was written; if this was done online, then you would need to go to the initial excel file (raw data file) and look up the ID to see what value was there.