2018-12-04
revised 2018-12-21

Project TIER Reproducibility Exercise: Animal House in Alcohol-Free Dorms?

I. Introduction

In this exercise, you will explore whether there is less drinking among college students who live in alcohol-free housing than among students who do not live in alcohol-free housing. The data you will use come from a survey of students at American colleges and universities conducted in 2001. The survey focused on alcohol use, but also included questions on many other aspects of college life.

This exercise asks you to use the data from the alcohol survey to create a number of bar graphs, and then answer a series of questions about what the bar graphs suggest about patterns of alcohol use among students that live in alcohol-free housing compared to those who do not.

A critical dimension of this exercise is documenting your work in such a way that everything you do with your data is transparent and reproducible. In this context, that means several things:

> Writing do-files containing code that executes every step of data processing and analysis you do to construct the bar graphs—including opening up the file containing the survey data, cleaning and processing the data as necessary, and then generating the bar graphs.
>
> Saving all your work—including your do-files, the original survey data file, and some accompanying documents—in a set of folders with a well-defined structure, and using relative directory paths in your do-files to indicate where files that need to be opened are located and where files that need to be stored should be saved.
>
> Constructing a Data Appendix that serves as a codebook or user's guide to your analysis dataset—the cleaned and processed version of the survey data that you use to generate the bar graphs.
>
> Writing a Read Me file that describes the content and organization of the documentation you create for this exercise, and gives step-by-step instructions for using the documentation to replicate all the data processing you did and to reproduce your bar graphs.

When you have completed this exercise, you will turn in both a printed report and a set of electronic files you create to document your work.

II. Create a set of folders to store your work in

Decide where you will keep your work. Depending on your preferences and the technology you are using, you may store these folders locally (on the hard disk of a computer you will be using), on a storage server or course management system maintained by your institution, or on-line [using a platform such as the Open Science Framework (www.osf.io), Dropbox (www.dropbox.com), or Google Drive (www.google.com/drive/)].

In any case, you should choose a place to store these folders that is secure and that you will be able to access easily while you work on this exercise.

Wherever you store these folders, be sure that there is system in place to automatically back them up, or that you manually maintain an up-to-date set of back-up files at all times.

Create your folders**.** An essential step toward ensuring that your data processing and analysis are easily reproducible is creating a folder hierarchy—a set of folders (and folders-within-folders) with a well-defined structure—in which you store your data, command files, and accompanying documents. You should create this set of folders at the very beginning of your work—before you even obtain the data you will be using.

You will keep all your files in these folders while you work on this exercise, and when you have finished you will use them to store the electronic documentation that you submit with your printed report.

To create the folder hierarchy for this exercise:

Make a new folder with the name **Your-Name-Alcohol-Exercise**, and save it in the location you chose to save your work in. This will be the main folder in which you store all of your work for this exercise.

Inside **Your-Name-Alcohol-Exercise**, create the following folders and sub-folders:

Original-Data (a sub-folder of **Your-Name-Alcohol-Exercise**)

Metadata (a sub-folder of **Original-Data**)

Command-Files (a sub-folder of **Your-Name-Alcohol-Exercise**)

Analysis-Data (a sub-folder of **Your-Name-Alcohol-Exercise**)

Graphs (a sub-folder of **Your-Name-Alcohol-Exercise**)

For-Data-Appendix (a sub-folder of **Graphs**)

For-Report (a sub-folder of **Graphs**)

Save this folder hierarchy in the location you chose to store your work for this exercise.

> ***Tip:*** Be sure to build this folder hierarchy before you proceed to the next steps of this exercise. Having these folders set up in advance so that you have a place to put the various files you obtain or create while you work on this exercise will help things go smoothly.

III. Download and examine the dataset and codebook you will use for this exercise

Find and explore the website where the data for this exercise are stored. The data used in this exercise come from a study titled *Harvard School of Public Health College Alcohol Study, 2001*. This study is archived at the Inter-University Consortium for Political and Social Research (ICPSR); its ICPSR study number is 4291.

Go to the ICPSR website (www.icpsr.umich.edu), and search for this study.

When you have found the page for this study, read the information provided there.

> ***Tip:*** Some of the information on the webpage for the dataset may have to do with technical matters you don't understand or details that may not appear important, but that is OK. Just read through the page and get out of it what you can.

Download the dataset and codebook. Before you can download the data, you need to create an account on the ICPSR website. Once you have done so, log in, go to the main page for the 2001 alcohol study, and click on the "Download" button. You will get a menu of formats to choose from, including SAS, SPSS, Stata, and ASCII. From this menu, select Stata.

The data file will be downloaded in a zipped folder. When you unzip and look inside the folder, you will see that it contains several documents, as well as a subfolder called **DS0001**. For this exercise, you will need to use only two of the files you downloaded, both of which will be stored in the **DS0001** subfolder:

04291-0001-Data.dta. This file contains the data from the survey, stored in Stata's *.dta* format. This file will be referred to as the "original data file" for this exercise.

04291-0001-Codebook.pdf. This is the codebook for the 2001 alcohol survey. It will be referred to as the "metadata" for the original data file.

You should save copies of both of these files in the folder hierarchy you created for this exercise.

Save the original data file in the **Original-Data** folder.

Save the codebook in the **Metadata** folder (which is inside your **Original-Data** folder).

Examine the codebook and dataset. Examine the codebook and original data file to familiarize yourself with the information they contain. In particular, by reading the codebook and using Stata to explore the contents of the data, answer the following questions:

What is the recommended format for citations of this dataset (to use, for example, in a reference list or a list of data sources)?

How many observations are there in the dataset?

How many variables are there in the dataset?

What is the unit of analysis for the dataset?

What population was the sample represented in this dataset taken from?

For this exercise, you will need to use six variables from alcohol study. In the original data file, these variables have the names *A6*, *B8*, *B9*, *C13*, *G6*, and *G11*. For each of these variables, answer the following questions:

How is the variable defined, or what does it represent?

What units is the variable measured in, or what is the coding scheme?

Is it categorical or quantitative? If it is quantitative, find the mean, standard deviation, the three quartiles, the minimum and the maximum. It if is categorical, how many categories are there, what do they represent, and what proportion of the observations fall in each category? For each categorical variable, decide whether the categories are ordered or unordered

Finally, suppose you wanted know what college or university each student in the sample attended. Would you be able to find that out using the codebook and dataset you downloaded from ICPSR? If so, how would you do so? If not, do you have any idea why that information is not provided?

> *Tip:* You will not need to turn in written answers to these questions, and you will not need to turn in do-files that document this preliminary exploration of your data. However, it is important to convince yourself that you understand all of the questions well. If you are hazy on any of them, you are likely to encounter trouble as you work on the parts of the exercise that you do have to turn in.

IV. General comments about organizing your work and writing do-files

Folder structure and file locations. Before you begin working on your do-files, make sure that you have created a folder hierarchy as described in section II, and that you have put the original data and the codebook in the folders specified section III. As you work on this exercise, you will be creating new files of various kinds, and storing them in designated locations in this hierarchy.

Whenever you are working with Stata, a copy of your complete folder hierarchy, with the most up-to-date versions of all your files stored in their designated locations, should be saved on the computer you are working on.

> *Tip:* Throughout this exercise, it is essential to keep the folder hierarchy, as well as the locations of your files within the folder hierarchy, fixed. In other words, don't rearrange your folders or move any of your files around.

Keeping track of the working directory: A simple convention. An essential part of making your data processing and analysis reproducible is keeping track of which folder is set as the working directory when Stata is running.

For this exercise, adopt the following convention:

Designate your **Command-Files** folder as Stata's working directory at all times.

In particular:

> Be sure that your **Command-Files** folder is designated as Stata's working directory while you are exploring the data and writing your do-files.
>
> Write your do-files under the assumption that the **Command-Files** folder will be designated as Stata's working directory when they are executed.

Whenever you begin a session of work, start by checking to see whether the **Command-Files** folder is initially designated as Stata's working directory; if not, change the working directory to **Command-Files**. After that, don't change the working directory again.

To make sure that anyone using your documentation understands that the **Command-Files** folder should be designated as the working directory, put a header at the beginning of every do-file reminding the user of that fact.

Relative directory paths. In the do-files you write for this exercise, you will sometimes need to write commands that refer to particular folders in your hierarchy: Commands that open up existing files need to specify the locations of the folders that contain the files to be opened, and commands that save files that have just been created need to specify the location of the folder in which the files should be stored.

Whenever you write a command in which you specify a folder location, you should do so using a relative directory path—i.e., the path to the folder starting from whatever folder is currently set as Stata's working directory. Given the convention recommended above, the working directory will be the **Command-Files** folder.

You should not specify the location of a folder with an absolute directory path—i.e., the path to the folder starting from the root directory (usually the C: drive) of the computer you are working on. Do-files containing commands that use absolute directory paths will not run on other computers that have different directory structures.

Comments. Throughout each of your do-files, insert comments that explain in words what each command or sequence of commands is doing.

> ***Tip:*** Putting detailed comments in your do-files is useful to anyone using your documentation to reproduce your work. But detailed comments will also be useful to you: you will be amazed at how obvious something may seem to you while you are working on it, yet how confusing it can be when you come back to it a few days later. When you write good comments as you go along, you won't have to struggle figure out what the commands you wrote in your do-files on previous occasions are all about.

V. Writing the do-files

Overview

For this exercise, you will write three do-files:

processing.do. In most studies involving statistical data (including this exercise), the original data must be processed before they are ready to be analyzed. For example, this processing may include correcting or removing errors in the original data, dropping variables or observations, generating new variables, and (when your original data are stored in more than one file) combining data files. Data on which all the processing necessary to prepare them for analysis has been completed are referred to as "analysis data." These are the data that will be used to generate the results that are presented in the final report on the study.

The *processing.do* do-file you write for this exercise will contain commands that open up the original data file (*04291-0001-Data.dta*), process the data as necessary to create your analysis data file, and then save the analysis data in a file named *analysis.dta*.

data-appendix.do. The Data Appendix is a document that serves as a codebook or user's guide for your analysis data. It contains variable definitions, coding and summary statistics (numeric and graphical) for all the variables in your analysis data.

The *data-appendix.do* do-file will contain commands that open up your analysis data file (*analysis.dta*) and then generate the summary statistics and other information that you will include in the Data Appendix.

results.do. Figures, tables and other information presented in a report about a study that were generated by computations performed on statistical data are referred to as "results."

For this exercise, the results consist of six bar graphs. The *results.do* do-file will contain commands that open up *analysis.dta*, and then generate those graphs, saving each graph when it is created.

Specific instructions

Instructions for writing these three do-files are given below.

> ***Tip:*** These instructions are detailed, but note that they are not a line-by-line map of the commands you must write. Instead, you need to figure out what individual command or sequence of commands you need to write to ensure that each of the required tasks is carried out.

At the beginning of every do-file you write, remember to write a header indicating that the working directory should be set to the **Command-Files** folder when the do-file is executed.

> *Tip:* Here's an example of what that header might say: "Stata's working directory should be set to the **Command-Files** folder whenever this do-file (or any part of it) is executed. Before running this do-file, check Stata's working directory; if necessary, change it to the **Command-Files** folder."

processing.do

In *processing.do*, you should write commands that accomplish the following:

--Open the *04291-0001-Data.dta* data file.

> *Tip:* As indicated above, this file should be stored in your **Original-Data** folder. Since that is not the folder that will be designated as Stata's working directory, the command you write that opens the original data file will need to specify the folder in which it is located. As always, the location of the folder should be specified with a relative directory path.

--For this exercise, we want to consider only students who live in campus housing or sororities/fraternities. Therefore, drop all cases for which the student's residence at college is "Off campus house/apt" or "other." Also, drop all cases for which the variable representing the student's residence at college has a missing value.

--Generate a variable called *drunk*, which is:

equal to 0 if the student drank enough to get drunk fewer than three times in the last thirty days

equal to 1 if the student drank enough to get drunk three or more times in the last thirty days

equal to missing (".") if the variable indicating how many times the student drank enough to get drunk in the last thirty days is missing

--Generate a variable called *hsdrunk*, which is:

equal to 0 if the student had five or more drinks in a row on two or fewer occasions during her/his last year of high school

equal to 1 if the student had five or more drinks in a row on three or more occasions during her/his last year of high school equal to missing (".") if the variable indicating how many times the student had five or more drinks in a row during her/his last year of high school is missing

--Generate a variable called *free*, which is:

equal to 0 if the student does not live in alcohol-free housing

equal to 1 if the student lives in alcohol-free housing

equal to missing (".") if the variable indicating whether or not the student lives in alcohol-free housing is missing

--Generate a variable called *volfree*, which is:

equal to 1 if *free*=1 (the student lives in alcohol-free housing) and the student requested to live in alcohol-free housing

equal to 0 if *free=1* (the student lives in alcohol-free housing) and the student was assigned to live in alcohol-free housing

equal to missing (".") in all other cases

--Generate a variable called *housing*, which is:

equal to 1 if *free*=0 (the student does not live in alcohol-free housing)

equal to 2 if *free*=1 (the student lives in alcohol-free housing) and the student was assigned to live in alcohol-free housing

equal to 3 if *free*=1 (the student lives in alcohol-free housing) and all on-campus housing is alcohol-free

equal to 4 if *free*=1 (the student lives in alcohol-free housing) and the student requested to live in alcohol-free housing

equal to missing (".") in all other cases

--Generate a new variable that is simply an exact copy of *G6*, and give it the name *health*.

Tip: You will not use the variable *health* in this exercise, but in the addendum to this document it is used to illustrate the information that should be provided in the Data Appendix.

--Drop all variables other than the six that were just created.

Tip: It will be easier to do this using Stata's `keep` command than Stata's `drop` command.

--Assign appropriate labels to each of the six variables.

> ***Tip:*** The variable names you defined when you created the variables (*drunk*, *hsdrunk*, etc.) are the ones you will use when you write commands for Stata. The labels you apply to the variables are the ones that will appear in figures and tables that Stata produces. Choose short but descriptive labels. For example, an appropriate label for *health* would be "Health Rating".

--Assign appropriate labels to the values of each of the six variables. For example, based on information given for the variable *G6* in the original data file, appropriate value labels for the variable *health* would be "Excellent" for the value 1, and "Very Good" for 2, "Good" for 3, "Fair" for 4, and "Poor" for 5.

--Drop all cases for which the value of *drunk* is missing.

--Drop all cases for which the value of *hsdrunk* is missing.

--Drop all cases for which the value of *housing* is missing.

--Save the new data file you have created. This will be your analysis data file—i.e., the fully processed data that you will use to generate the bar graphs for this exercise. Give this analysis data file the name *analysis.dta*, and save it in your **Analysis-Data** folder. (Use a relative directory path to specify the location of the **Analysis-Data** folder.)

Remember to save *processing.do* in your **Command-Files** folder.

<u>data-appendix.do</u>

The *data-appendix.do* do-file will contain commands that generate information that will appear in the Data Appendix (see section VII below). In particular, *data-appendix.do* should contain commands that accomplish the following:

--Open the data file *analysis.dta* (which, after you have written and run *processing.do*, should be saved in your **Analysis-Data** folder).

--Generate the following output for each of the five variables (excluding *health*) in the *analysis.dta* data file:

--The number of missing observations and the total (missing plus non-missing) number of observations.

--The frequency distribution and the percent frequency distribution.

--A bar graph showing the percent frequency distribution. When each bar graph is created, a copy should be saved (as a Stata *.gph* graphics file) in the **For-Data-Appendix** subfolder of the **Graphs** folder. (In the commands that save these graph files, use relative directory paths to specify the location of the **For-Data-Appendix** sub-folder.) Save each graph with the name `*var'-dist.gph*, where `*var'* is replaced by the name of the variable.

Write comments explaining which commands produce which information and graphs: Each command that produces information or a graph that will be included in your Data Appendix should be preceded by a comment that describes the output the command will produce. For example, just above a command that generates a table showing the frequency distribution of the variable *housing*, you would write a comment like "The following command generates a table showing the frequency distribution of "*housing*"."

Remember to save *data-appendix.do* in your **Command-Files** folder.

analysis.do

The *analysis.do* do-file generates the bar graphs that constitute the main analysis you do for this assignment, using the data in your final *analysis.dta* file.

Write commands in *analysis.do* that accomplish the following:

--Open the data file *analysis.dta* (which, after you have written and run *processing.do*, should be saved in your **Analysis-Data** folder).

--Generate a bar graph, with two bars, where:

one bar represents students living in alcohol-free housing

one bar represents students who do not live in alcohol-free housing

the height of each bar is equal to the proportion of students (in the group the bar represents) who drank enough to get drunk three or more times in the last thirty days

Save this graph in the **For-Report** subfolder of the **Graphs** folder, with the name *Figure1.gph*. (Use a relative directory path to specify the location of the **For-Report** sub-folder.)

--Generate a bar graph, with two bars, where:

one bar represents students who live in alcohol-free housing because they requested it

one bar represents students who live alcohol free housing because they were assigned to it the height of each bar is equal to the proportion of students (in the group the bar represents) who drank enough to get drunk three or more times in the last thirty days

Save this graph in the **For-Report** subfolder of the **Graphs** folder, with the name *Figure2.gph*.

--Generate a bar graph, with four bars, where:

each bar represents students in one of the four categories defined by the variable *housing*

the height of each bar is equal to the proportion of students (in the group the bar represents) who drank enough to get drunk three or more times in the last thirty days

Save this graph in the **For-Report** subfolder of the **Graphs** folder, with the name *Figure3.gph*.

--Generate two side-by-side bar graphs, each consisting of two bars, where:

one of the bar graphs represents only students who had five or more drinks in a row on three or more occasions during their last year of high school

one of the bar graphs represents only students who had five or more drinks in a row on fewer than three occasions during their last year of high school

and in each of these graphs,

one bar represents students who live in alcohol-free housing

one bar represents students who do not live in alcohol-free housing

the height of each bar is equal to the proportion of students (in the group the bar represents) who drank enough to get drunk three or more times in the last thirty days

Save this graph in the **For-Report** subfolder of the **Graphs** folder, with the name *Figure4.gph*.

--Generate two side-by side bar graphs, each consisting of two bars, where:

one of the bar graphs represents only students who had five or more drinks in a row on three or more occasions during their last year of high school

one of the bar graphs represents only students who had five or more drinks in a row on fewer than three occasions during their last year of high school

and in each of these graphs,

one bar represents students who live in alcohol-free housing because they requested it

one bar represents students who live alcohol free housing because they were assigned to it

the height of each bar is equal to the proportion of students (in the group the bar represents) who drank enough to get drunk three or more times in the last thirty days

Save this graph in the **For-Report** subfolder of the **Graphs** folder, with the name *Figure5.gph*.

--Generate two side-by side bar graphs, each consisting of four bars, where:

one of the bar graphs represents only students who had five or more drinks in a row on three or more occasions during their last year of high school

one of the bar graphs represents only students who had five or more drinks in a row on fewer than three occasions during their last year of high school

and in each of these graphs,

each bar represents students in one of the four categories defined by the variable *housing*

the height of each bar is equal to the proportion of students (in the group the bar represents) who drank enough to get drunk three or more times in the last thirty days

Save this graph in the **For-Report** subfolder of the **Graphs** folder, with the name *Figure6.gph*.

Write comments indicating which commands produce which graphs: For example, before the command that generates Figure 4, write a comment like “The following command generates Figure 4.”

Remember to save *analysis.do* in your **Command-Files** folder.

Cleaning up your do-files

Writing the do-files that accomplish the tasks outlined above will be a large part of the work you do for this exercise. For most of the tasks, your first attempt at writing a command or commands won't work—you will get error messages, or something will be obviously incorrect, and you will need to try again, maybe several times, before you figure out syntax that works. When you have finally gotten everything in a do-file to run successfully, you may well have accumulated a lot of extra stuff that is no longer relevant: commands that were first tries, false starts, or exploration to figure out how things work.

At the end of the exercise, when you are preparing the final electronic documentation to submit with your printed report, you should clean up your do-files: remove any first tries, false starts and exploration that turned out to be mistakes or that do not serve any purpose. The only commands left in the do-files you submit with your documentation should be ones that do something necessary to accomplish the required tasks.

And remember to include comments throughout your do-files explaining the purpose of the commands at each step of the data processing and analysis.

> *Tip:* Your do-files of course need to run properly and execute the steps of data processing and analysis that reproduce your results. But they should also be useful documents to a human reader who wants to learn about what you did for this project. Keep that goal in mind as you do your final clean-up of the commands and insert comments.

VI. The Read Me file

The Read Me file is a document that provides information about the electronic files you have assembled to document your work for this exercise. It should consist of two main sections.

The first section should provide an outline or map of all the files included in the replication documentation, and the folders and sub-folders in which they are stored.

The second section should give step-by-step instructions explaining how to use your documentation to (i) replicate all the data processing required to transform the original data into your analysis data file, (ii) generate all the computer output that you used in the Data Appendix, and (iii) reproduce the six bar graphs you created.

Those instructions should be written in plain English, and they should be clear and detailed enough that someone unfamiliar with this exercise would actually be able to follow them and replicate everything you did with your data.

Formatting the Read Me file

The title "Read Me File for XXX" (where you substitute an informative title for XXX) should appear at the top of the document.

The name(s) of the authors(s) of the exercise and the date it was turned in should also be shown at the top of the Read Me file.   The pages of the Read Me file should be numbered.

When the Read Me file is complete, you should save a copy of it in *pdf* format. Give the file the name *read-me.pdf*.

A copy of *read-me.pdf* should be stored in the top level of your main project folder (**Your-Name-Alcohol-Exercise**).

VII. The Printed Report

The printed report you turn in for this exercise will consist of three parts: answers to a series of questions that ask you to interpret the bar graphs you created, the Data Appendix, and printed copies of the graphs.

Interpretation

This section of the report should begin with the header "**I. Interpretation**." In the body of this section, you should give answers to the following questions:

1) Describe the sample in the *analysis.dta* data file you created. What is the unit of observation? How many observations are there? Describe the method used to create the sample, and any additional criteria used to include or exclude any groups of individuals.

Note that this question is about the cleaned and processed data in the *analysis.dta* file, rather than the original dataset you downloaded from ICPSR. Nonetheless, to answer it you will need to refer to the codebook for the original study. However, you will also need to take into account the processing of the data that you did to create *analysis.dta*.

2) Explain in words what Figure 1 shows. Is that what you would have expected? If not, how does Figure 1 differ from what you expected?

3) Explain in words what Figure 2 shows. Does what you see in Figure 2 shed any light on what you saw in Figure 1? Explain.

4) How many individuals are represented in Figure 1? (That is, how many observations did Stata use to generate the figure?) And how many individuals are represented in Figure 2? If the numbers of individuals represented in Figures 1 and 2 are not the same, explain: were some individuals represented in Figure 1 not represented in Figure 2? If so, which individuals were these? And were some individuals represented in Figure 2 not represented in Figure 1? If so, which individuals were these?

5) Consider Figure 3. Which parts of it show information that was already shown in Figures 1 and 2? Which parts of it show new information? Summarize in words what Figure 3 shows.

6) Compare Figure 4 to Figure 1. Explain what new information is provided by Figure 4, and give some interpretation.

7) Compare Figure 5 to Figure 2. Explain what new information is provided by Figure 4, and give some interpretation.

8) Compare Figure 6 to Figure 3. Explain what new information is provided by Figure 4, and give some interpretation.

9) From the figures you created, what general conclusions can you draw about the factors associated with drinking among college students?

The Data Appendix

This section should begin with the heading "**II. Data Appendix**."

The Data Appendix should consist of multiple sections, or “entries,” one for each of the variables in the final data set.

For each variable, the entry in the Data Appendix should give:[1]

--The name of the variable.

--The number of missing observations, in the format *m/n*, where *m* represents the number of missing observations and *n* represents the total number of observations (missing plus non-missing).

--A definition of the variable.

--The possible values of the variable.

--The coding of the values; i.e., an explanation of what each value of the variable represents.

--A table showing the frequency distribution and the percent frequency distribution. If the categorical variable is an ordered categorical variable, this table should also show the cumulative percent frequency distribution.

[1] In general, the information presented in a Data Appendix for a particular variable depends on whether it is quantitative or categorical. In this assignment, however, all of the variables are categorical, so we won’t worry about what you present in the case of a quantitative variable. But if you are curious, see https://www.projecttier.org/tier-protocol/specifications/#the-data-appendix.

Tip: If the categorical variable is not an ordered categorical variable, be sure that this table does not show a cumulative percent frequency distribution.

--A bar graph that illustrates the percent frequency distribution.

To illustrate, an addendum to this exercise shows what the entry for the variable *health* would consist of. The Data Appendix you prepare for this assignment should consist of similar entries for the five other variables in the analysis data set. (You do not need to include an entry for the variable *health*.)

Figures

This section should begin with the header "**III. Figures**."

This section should show Figures 1-6, in order.

Write a caption below each figure that indicates which figure number it is, and gives an informative title for the figure. For instance, an appropriate caption for Figure 1 would be "Figure 1: Rates of Heavy Drinking in Alcohol-Free vs. Not Alcohol-Free Housing."

TIP: Be sure that each caption appears on the same page as the figure it describes.

Formatting of the printed report:

The document should begin with a cover page that gives:

--an informative title (chosen by you)

--the name(s) of the author(s) of the document

--the date the assignment was turned in

The cover page should be followed by the three sections of content described above, each beginning with the appropriate header.

The pages of the document should be numbered.

Save an electronic copy of the report:

When the report is complete, you should save a copy of it in *pdf* format. Give the file the name *report.pdf*.

Tip: report.pdf should be a single electronic document that contains all parts of the printed report: the cover page and all three content sections.

A copy of *report.pdf* should be stored in the top level of your main project folder (**Your-Name-Alcohol-Exercise**).

VIII. Electronic Documentation

When you have completed the exercise, you should be sure that clean, final versions of all the electronic documents you created are stored in the appropriate locations in the folder hierarchy you created for this exercise.

Here is a map of which documents should be in which folders:

Your-Name-Alcohol-Exercise (the main project folder)

readme.pdf (a document in the top level of **Your-Name-Alcohol-Exercise**)

report.pdf (a document in the top level of **Your-Name-Alcohol-Exercise**)

Original-Data (a sub-folder of **Your-Name-Alcohol-Exercise**)

04291-0001-Data.dta

Metadata (a sub-folder of **Original-Data**)

04291-0001-Codebook.pdf

Command-Files (a sub-folder of **Your-Name-Alcohol-Exercise**)

processing.do

data-appendix.do

analysis.do

Analysis-Data (a sub-folder of **Your-Name-Alcohol-Exercise**)

analysis.dta

Graphs (a sub-folder of **Your-Name-Alcohol-Exercise**)

For-Report (a sub-folder of **Graphs**)

figure1.gph

figure2.gph

figure3.gph

figure4.gph

figure5.gph

figure6.gph

For-Data-Appendix (a sub-folder of **Graphs**)

drunk-dist.gph

free-dist.gph

volfree-dist.gph

housing-dist.gph

hsdrunk-dist.gph

Don't leave any junk lying around in these folders.

Make sure your folders don't include any sub-folders other than the ones shown above.

Make sure your folders don't contain any files other than the ones shown above.

Test your electronic files to make sure they run.

Find a computer that is not one you have been using while working on this exercise.

Copy your entire project folder (**Your-Name-Alcohol-Exercise**), with all its contents, onto that computer.

Then be sure that you can complete the following steps without needing to move around any of the folders or files:

Launch Stata.

Set the working directory to your **Command-Files** folder.

Run the *processing.do* do-file. After you run this file, check to be sure that the *analysis.dta* data file that you had stored in your **Analysis-Data** folder was overwritten by a new copy that was created when you ran this do-file. (You can check that by looking at the "date modified" information for the file.)

Run the *data-appendix.do* do-file. Verify that running this do-file generates the desired output. Also check to be sure that the graphs that were stored in the **For-Data-Appendix** subfolder of the **Graphs** folder were overwritten by new versions that were generated when you ran this do-file.

Run the *analysis.do* do-file. After you run this file, check to be sure that the six graph files that were stored in the **For-Report** subfolder of the **Graphs** folder were overwritten by new copies.

Tip: If you encounter any hitches, do not submit your documentation for this exercise until you have figured out what the problems are and resolved them.

ADDENDUM
An Example of an Entry in the Data Appendix

This addendum uses the variable *health* (which you did not use for this exercise, but is included in the *analysis.dta* data file) to illustrate the format to follow in the Data Appendix.

Variable name: *health*

Missing observations: 4/2717

Definition: “Respondent’s answer to the question: ‘In general, how would you rate your health now?’” (The wording of this question was taken *verbatim* from item G6 in the questionnaire used for the survey; the questionnaire is included in the codebook for the study.)

Possible values: 1, 2, 3, 4, 5.

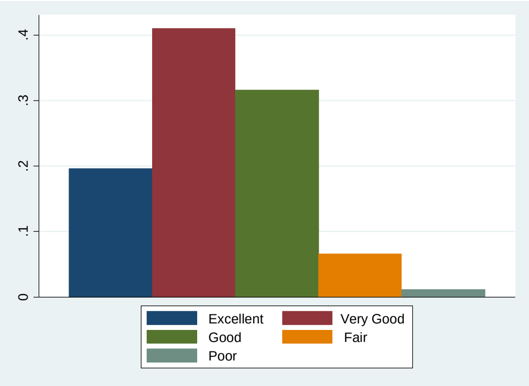
Coding:

Value	Meaning
1	Excellent
2	Very Good
3	Good
4	Fair
5	Poor

Frequency table:

```
Health Rating |      Freq.     Percent        Cum.
---------------+-----------------------------------
  Excellent    |        532       19.61       19.61
  Very Good    |      1,113       41.02       60.63
       Good    |        858       31.63       92.26
       Fair    |        179        6.60       98.86
       Poor    |         31        1.14      100.00
---------------+-----------------------------------
      Total |      2,713      100.00
```

Bar graph showing percent frequency distribution:


Excellent
Very Good
Good
Fair
Poor