Statistics Package Exercise: Exploring Variables in a Dataset

Learning Objective: Classify a data analysis situation (involving two variables) according to the "role-type classification," and state the appropriate display and/or numerical measures that should be used in order to summarize the data.

Let's Explore a Dataset

In this activity we

- Learn how to open and examine a dataset.
- Practice classifying variables by their type: quantitative or categorical.
- Learn how to handle categorical variables whose values are numerically coded.

Background to Dataset

Clinical depression is the most common mental illness in the United States, affecting 19 million adults each year (Source: NIMH, 1999). Nearly 50% of individuals who experience a major episode will have a recurrence within 2 to 3 years. Researchers are interested in comparing therapeutic solutions that could delay or reduce the incidence of recurrence.

In a study conducted by the National Institutes of Health, 109 clinically depressed patients were separated into three groups, and each group was given one of two active drugs (imipramine or lithium) or no drug at all. For each patient, the dataset contains the treatment used, the outcome of the treatment, and several other interesting characteristics.

Here is a summary of the variables in our dataset:

- **Hospt:** The patient's hospital, represented by a code for each of the 5 hospitals (1, 2, 3, 5, or 6)
- **Treat:** The treatment received by the patient (Lithium, Imipramine, or Placebo)
- **Outcome:** Whether or not a recurrence occurred during the patient's treatment (Recurrence or No Recurrence)
- **Time:** Either the time in days till the first recurrence, or if a recurrence did not occur, the length in days of the patient's participation in the study.
- **AcuteT:** The time in days that the patient was depressed prior to the study.
- **Age:** The age of the patient in years, when the patient entered the study.
- **Gender:** The patient's gender (1 = Female, 2 = Male)

- R | StatCrunch | TI Calculator | Minitab | Excel

R Instructions
To open R (a free software environment for statistical computing and graphics) with the data preloaded, right-click here and choose "Save Target As" to download the file to your computer. Then find the downloaded file and double-click it to open it in R.

The data have been loaded into the data frame depression. Enter the command

```
depression
```

to see the data.

Note: Using R—Throughout the statistics package exercises in this course, you will be given commands to execute in R. If you type them in by hand be aware that R is sensitive to capitalization, spelling, and format. After you type a command into the R console press <Enter> to execute the command. You can use the following steps to avoid having to type all of these commands in by hand:

1. Highlight the command with your mouse.
2. On the browser menu, click "Edit," then "Copy."
3. Click on the R command window, then at the top of the R window, click "Edit," then "Paste."
4. You may have to press <Enter> to execute the command.

When you enter the command

```
depression
```

into R, you will see a large data table. Each row of this table contains the values of the variables associated with a single individual, and the different variables are separated into columns. The columns are labeled with the variable names

columns are labeled with the variable names.

If we simply wanted to observe only

```
Age
```

information we can extract that specific variable from the data frame by connecting the data frame name to the column name using the

```
$
```

symbol, such as in the following command:

```
depression$Age
```

Often it is easier to use labels for categorical variables that are as close as possible to the meanings of the categories. Now we will recode the variable gender with the labels "Male" and "Female." Copy the entire following command into R.

```
depression$Gender =
replace(depression$Gender,depression$Gender==1,'Female');
```

```
depression$Gender =
replace(depression$Gender,depression$Gender==2,'Male');
```

```
depression$Gender
```

Remember, you may have to press <Enter> to execute the command.

Notice that the column Gender now contains the meaningful labels "Female" and "Male" where before it contained "1" and "2" codes.

Note: Using R - To learn more about any command names you see in these notes, enter

```
help(commandname)
```

or into R or check out the resources listed under the Help menu.

Did I Get This (1/1 point)

What are the categorical variables in this dataset?

Your Answer:

hospt, treat, outcome, gender

Our Answer:

The categorical variables are 1) Hostp because the numbers represent codes, which are used to identify individual hospitals and place them into categories. As such, the numbers used for the codes (1, 2, 3, 5, and 6) have no arithmetic meaning; 2) Treat because the treatment received by the patients is in the form of categories (Lithium, Imipramine, or Placebo); 3) Outcome since recurrence is in the form of two categories (Recurrence or No Recurrence) and 4) Gender because the numbers represent two distinct categories: Female and Male. Thus, the numbers used to represent gender (1 = Female; 2 = Male) have no arithmetic meaning.

Resubmit Reset

Did I Get This (1/1 point)

What are the quantitative variables in this dataset?

Your Answer:

time, acutet, age

Our Answer:

The quantitative variables are 1) Time since it can take on multiple numerical values, which have arithmetic meaning (i.e., it makes sense to add, subtract, multiply, divide, or compare the magnitude of such values); 2) Age since it can take on multiple numerical values, which represent a characteristic of the patient; and 3) AcuteT because it can take on multiple numerical values to represent a characteristic of the patient.

Resubmit Reset