Describing Data in R

To see what the type of data each column is stored is, we need to find out what structure each is saved as. To do this we use the str() function.

```
str(sleep)
```

This uses a function. For help with functions see the help guide ”Functions in R”. First we need to say str as it is the name of the function. We need sleep as we need to specify the dataframe . We're telling the function to look in the dataframe sleep and find the structure of each column. You should get the following output:

```
  'data.frame': 20 obs. of  3 variables:
 $ extra: num  0.7  1.6  0.2  1.2  0.1 3.4 3.7 0.8 0 2 ...
 $ group: Factor w/ 2 levels ”1”,”2”: 1 1 1 1 1 1 1 1 1 1 ...
 $ ID   : Factor w/ 10 levels ”1”,”2”,”3”,”4”,..: 1 2 3 4 5 6 7 8 9 10 ...
```

First we get a line that confirms that sleep is, in fact, a data frame. Then we are told that this dataframe has three variables (columns) with 20 observations (rows) of each variable. Then it goes through and lists each column of variable name stored in this dataframe and tells us what type of variable it as stored as. For example, we can extra is a numeric variable because it says num after it. group and ID are factor level variables. We can see that group has two different options, either 1 or 2. This corresponds to the different levels of group.

3 Describing Continuous variables

Our first step in this guide will be to describe the continuous variable extra in the data frame sleep. We'll start by taking the mean and standard deviation for extra variable as a whole, and then we will find the mean and standard deviation of the extra variable for each of the intervention groups 1 and 2.

The mean and sd function can be used to find the mean and standard deviation of the extra function as follows:

```
mean(sleep$extra)
```

```
sd(sleep$extra)
```

Here the mean or the sd function tells R we want to calculate the mean or the standard deviation respectively. The sleep$extra says that we want R to find the sleep dataframe and select the extra column. You should have found the mean is 1.54 and the sd is 2.02.

Now we can use those same functions to calculate the mean and standard deviation of each group. To do this we need to be able to tell R to find the data frame sleep and find the column sleep but only return the elements of sleep that correspond to the column group equalling ”1” or ”2”. The first line of code below does this for the first group followed by a line to do this for the second group. The square brackets are used to tell R that we will be specifying which of the sleep$extra numbers to keep, and bit inside the square brackets sleep$group==”1” tells R to look for where the group variable is equal to 1.

```
sleep$extra[sleep$group==”1”]
```

```
sleep$extra[sleep$group==”2”]
```

extra
function as follows:

```
sd(sleep$extra[sleep$group=="2"])
```

In this section we found the mean and standard deviation of a continuous variable, and then the mean and variances of a grouping variable broken into two different conditions.

4 Describing Grouping Variables

In the previous section we looked at how we might summarise a continuous variable by taking its