Demo 2: Preparing data for statistical analysis

Experimental and Statistical Methods in Biological Sciences I

September 15, 2014

In these exercises, you will practice the steps taken when preparing your data for statistical analyses. We will get familiar with the 'naming' data set.

1 Load in the data

Start R.

Start by creating a new script where you save all your work for future purposes. Remember to add comments (with #) so that you and others will be able to read the script later.

Navigate to the correct directory.

The data can be found at `http://bit.ly/RotCjb` . Load it into a data frame called `naming` and find out what's in it.

```
> naming <- read.csv("http://bit.ly/RotCjb", header=TRUE, sep="\t")
```

Question 1: Using the built-in help, make sure you understand why we had to use the arguments as above. Try `?read.csv`.

Question 2: What are the columns called? Try `names`.

Question 3: Describe how the data looks like.

2 Factors

Question 4: Use the `summary` function to look at `naming$hrs` and `naming$word.type`. Explain why the two summaries are different.

Question 5: What kind of objects are each of `naming$hrs` and `naming$word.type`? Try `class` and `summary`.

Question 6: Take a closer look at the whole dataset by using `summary` for the whole data frame. Given what you know about factors, are any of the summaries surprising? Why?

Question 7: Change the anomalous column to a factor.

Tip: You will need to assign the 'factorized' value of the column to itself.

Question 8: Add meaningful textual labels to this factor. Make this new factor into a new variable `naming$sex`. Remove the old one.

Tip: You will need to do some subsetting for the exercise; you can either use logical subscripting, or you may want to find out about the `subset` function.

Tip: To make things easier, use `attach` so that you don't have to type the long version all the time. You should do this when the names of your variables don't change anymore.

3 Missing values

Question 9: Now, take a closer look at the various summaries of `naming`. Do you find anything peculiar? Why?

Question 10: Locate the suspect values of `naming`. Remove these by setting them to a 'missing value'. How many entries are missing?

Question 11: How many participants are included in the complete dataset? (Careful!) To get the whole story, use `is.na`.

4 Sampling

For practising purposes...

Question 12: Add a new column to the data frame which numbers each of the participants.

Tip: `rep` and the `:` operator are your friends here, but remember there's no numeric relationship between participants... (i.e., make it a factor!)

Question 13: Use the `sample` command to make a list of 100 unique numbers between 1 and 1000.

Question 14: Using this list, and subscripting, make a new data frame which includes the data from 100 randomly-selected participants.

Tip: Remember that you want all the columns for specific rows of the data.

Question 15: Examine the data. Does it differ from the original data?

5 Finally...

After playing around with the data, write a description of what the naming data contains. Make a table describing the variables and their data types. Describe how the underlying study could be like. What kind of hypotheses/questions could you construct with these variables?

Remember to save your script and the data frames you created.